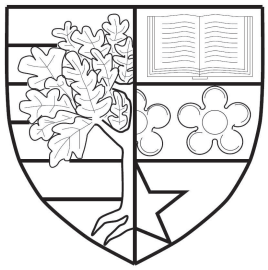
SENTIMENT ANALYSIS FOR MICRO-BLOGGING PLATFORMS IN ARABIC

by

Eshrag Ali Ahmad Refaee



Submitted for the degree of
Doctor of Philosophy

Department of Computer Science
School of Mathematical and Computer Sciences
Heriot-Watt University

July 2016

Abstract

Sentiment Analysis (SA) concerns the automatic extraction and classification of sentiments conveyed in a given text, i.e. labelling a text instance as positive, negative or neutral. SA research has attracted increasing interest in the past few years due to its numerous real-world applications. The recent interest in SA is also fuelled by the growing popularity of social media platforms (e.g. Twitter), as they provide large amounts of freely available and highly subjective content that can be readily crawled.

Most previous SA work has focused on English with considerable success. In this work, we focus on studying SA in Arabic, as a less-resourced language. This work reports on a wide set of investigations for SA in Arabic tweets, systematically comparing three existing approaches that have been shown successful in English. Specifically, we report experiments evaluating fully-supervised-based (SL), distant-supervision-based (DS), and machine-translation-based (MT) approaches for SA. The investigations cover training SA models on manually-labelled (i.e. in SL methods) and automatically-labelled (i.e. in DS methods) data-sets. In addition, we explored an MT-based approach that utilises existing off-the-shelf SA systems for English with no need for training data, assessing the impact of translation errors on the performance of SA models, which has not been previously addressed for Arabic tweets. Unlike previous work, we benchmark the trained models against an independent test-set of >3.5k instances collected at different points in time to account for topic-shifts issues in the Twitter stream. Despite the challenging noisy medium of Twitter and the mixture use of Dialectal and Standard forms of Arabic, we show that our SA systems are able to attain performance scores on Arabic tweets that are comparable to the state-of-the-art SA systems for English tweets.

The thesis also investigates the role of a wide set of features, including syntactic, semantic, morphological, language-style and Twitter-specific features. We introduce a set of affective-cues/social-signals features that capture information about the presence of contextual cues (e.g. prayers, laughter, etc.) to correlate them with the sentiment conveyed in an instance. Our investigations reveal a generally positive impact for utilising these features for SA in Arabic. Specifically, we show that a rich set of morphological features, which has not been previously used, extracted using a publicly-available morphological analyser for Arabic can significantly improve the performance of SA classifiers. We also demonstrate the usefulness of language-independent features (e.g. Twitter-specific) for SA. Our feature-sets outperform results reported in previous work on a previously built data-set.

Dedication

To the soul of my father.

Acknowledgements

I would like to thank my supervisor, Verena Rieser, for her patient guidance, endless support and encouragement throughout my Ph.D. Verena has been a supervisor and a best friend. Her good advice and acts of kindness make her what it means to be the best person one can aim to become and to know, learn from and work with. I am especially thankful to Verena for her support and genuine empathy during the difficult times when I lost my father half way through my Ph.D. I would never have finished this thesis without her support.

I would like to thank Helen Hastie and Rob Pooley for the helpful comments and discussions in Verena's maternity leave. I thank members of the Interaction Lab from whom I have come to learn new things. I would like also to thank IT Helpdesk staff, especially Iain Mccrone. Admin staff in MACS department have been very helpful, especially Sandra McArthur, Claire Porter and Christine McBride.

I am very grateful to my mother and siblings, for their love and support. I would like to express my deep gratitude for the generous scholarship to pursue my postgraduate studies granted by the Royal Embassy of Saudi Arabia and the Cultural Bureau in London. Finally, I would like to extend thanks to Jazan University in Saudi Arabia for offering me a job as a Research Associate at the School of Computer Sciences.

بسم الله الرحمن الرحيم

إهداء

فَاحمد الّله و اثني عليه .. الحمد لّله الذي بنعمته تّتم الصَالحَات ... الحمد لّله حمدَا حمدَا .. اللّهم اجعل هذَا العمل خَالصَا متقبلَا لوجهك الكريم ... اللّهم لك الحمد لَا احصي ثنَاءً عليك انت كمَا اثنيت علي نفسك .. اللّهم لك الحمد كمَا ينّبغي لجلَال وجهك وعظيم سلطَانك.

امَا بعد، فَاني اهدي هذَا العمل الي:

روح ابي الغَالي الذي لم يري هذَا اليوم .. الي من دعَاني دكتوره منذ كنت في الصف الثَاني الابتدَاءي .. الي من لم يسمح لي القدر بودَاعه .. الي من قَال كل مَا رآني: مرحبَا بَالربيع في آذَار .... و ب (اشرَاق) بهجه الأنوَار... الَاستَاذ وَالَاديب وَالشَاعر ... ابي .. علي احمد الرفَاعي.

الي امي الحبيبه .. الي من حملت حقيبتي الي المدرسه يومَاً .. الي من سهرت بجوَاري ليَالي طويله .. الي من اعتصرت المَاً في غيَابي .. الي من اضَاءت دعوَاتهَا خطوَاتي .. الي امي .. امزينه محمد عز الدين.

الي اخوَاني الَاعزَاء .. جبرَان وزيد وتركي .. الي اخوَاتي الغَاليَات .. افرَاح وَايمَان وتوَامي اخلَاص .. شكرَاً لَايمَانكم بي .. شكرَاً لثقتكم الغَاليه التي حملتهَا معي كل يوم .. ليس هنَاك كلمَات شكر تفيكم حقكم فأنتم مصدر فرحي وَاملي .. وبحبكم ودعمكم اللَامتنَاهي (بعد الّله) صمدت طيله هذه السنوَات في الغربه ..

الي ولدي الذي لم انجبه ... ايَاد الرفَاعي.

وختَامَاً، اشكر منسوبي الملحقيه الثقَافيه السعوديه بلندن ومنسوبي جَامعه جَازَان لدعمهم طوَال فتره ابتعَاثي في المّملكه المتحده.

Social Media is the elephant in the room - no decision management system will escape the impact of social media. Social Media Monitoring, including sentiment analysis, will become more and more a commodity and focus will be on integration with decision-based systems.

— Olivier Jouve IBM

Contents

List of Tables

List of Figures

Acronyms

ARFF Attribute Relation File Format

BOW Bag Of Words

CV Cross-Validation

DA Dialectal Arabic

DS Distant Supervision

GS Gold-Standard

IG Information Gain

KNN K Nearest Neighbour

LDC Linguistics Data Consortium

ML Machine Learning

MT Machine Translation

MSA Modern Standard Arabic

NB Naïve Bayes

NER Named Entity Recognition

NLP Natural Language Processing

OOV Out-of-Vocabulary

POS Part Of Speech

SA Sentiment Analysis

SI Sentiment Intensity

SL Supervised Learning

SM Social Media

SVM Support Vector Machines

TC Text Classification

χ^2 Chi Square Statistics

Chapter 1

Introduction

1.1 Motivation

Over the past decade, there has been a growing interest in collecting, processing and analysing user-generated text from social media. As a sub-task of Affective Computing, Sentiment Analysis (SA) provides the means to mine the web automatically and summarise vast amounts of user-generated text into the sentiments they convey. That is, SA can be cast as a text classification problem wherein the task is to classify the personal attitudes conveyed in the text and determine the polarity of a given text utterance with respect to the author's perspective as being positive, negative or neutral.

The growth of research in automatic analysis of people's attitudes and sentiments has coincided with the increasing popularity of social media [112]. This is due to the fact that social media has made it possible for users from different backgrounds, languages and cultures to share their views, stances, attitudes and sentiments towards a wide spectrum of topics/entities/aspects [127]. For instance, social media platforms have provided their users with an opportunity to discuss points of agreement, and/or conflict, and hence, encourage rich debate about economic, social, cultural and political stances. This is where the research area of SA plays a major role in capturing and analysing the subjective content from text produced by the general public on social media [108].

The ability to classify sentiments is important to understand attitudes, opinions,

evaluations and emotions communicated among users across the world about current issues - answering the question of *'what is going on'*. In this context, world-leading organisations like Google and Microsoft have established their internal systems to carry out sentiment-analysis-related tasks [112]. The unprecedented volume and variety of highly opinionated social media content provide new opportunities for research on SA to serve a range of real-world applications, such as:

- assessing brand/product success [51, 27]
- anticipating stock market trends, and financial performance [40]
- detecting radical/extreme/suicide trends [1, 115],
- detecting public mood/national happiness [41, 110, 58],
- assessing the popularity of a political party/candidate, which involves political predictions of election outcomes [128, 168, 117, 114],
- as an input for disaster response systems (e.g. early warnings and identification of fire events) [133].

Despite being commercial, social or political, it can be deduced that the main goal of SA is (more or less) to support decision making. For instance, SA can aid learning about customers' perception of a certain product/service and hence, improve marketing plans [27]. Therefore, the availability of reliable SA systems is of great value to meet such a growing demand.

Being a rich resource of subjective text that conveys personal stances, Twitter is a valuable resource for research on SA to explore how people react to various topics [112]. Twitter, among other social networks, has been witnessing a flurry of novel research and become a target for large research projects. This involves exploiting content from Twitter to infer valuable knowledge from tweets including sentiments [177, 81, 123]. SA research on tweets is not only motivated by the vast amount of freely available data to crawl [135], but also by the popularity of Twitter. [1] Dodds et al. [58] state that the selection of Twitter and other sources of big data is motivated

[1] Twitter is ranked as the 10th most popular website in the world [161].

by the growing interest to study content of social networks due to their influence both at social and individual levels. In addition, Zaidan and Callison-Burch [184] point out the significance of Twitter in particular as a valuable resource with regard to the recent unstable political and social circumstances in the Middle East. In particular, analysing sentiments conveyed via social media can be of great impact as they have shown to be a key influencer on reshaping social and political systems, such as those in some Arab countries [112]. In [89, 90], the authors study sentiments conveyed in social media posts during the time of the Arabic Spring. They argue that the findings of such studies are important not only for the individual people, but also for political decision makers [44].

Arabic is amongst the top 10 most popular languages in Twitter (and in social web) [178, 58], with nearly 17M Arabic tweets per day [24]. Despite that, there has been limited work on SA of Arabic tweets in comparison with English. A possible reason is that most of the previous work on Arabic Natural Language Processing (NLP), including SA, has focused on developing NLP resources for Modern Standard Arabic (MSA), e.g. news corpora [7, 181]. Arabic tweets are typically written in one of the Arabic dialects, which differ substantially from MSA [88]. Therefore, the lack of resources available for SA of Arabic social media content has resulted in narrowing down research exploring this area. English, as a well-resourced language, has received a considerable amount of research for SA in social media. For instance, in 2013, 2014 and 2015, a series of shared-tasks, known as SemEval, were conducted to carry out evaluations for SA systems. SemEval's series involve a popular task for SA on English tweets [123, 146, 145]. Such competition is valuable to encourage participants from all-over the world to develop and compare SA systems. Besides the comparative findings produced, research on English SA has also benefited from linguistic resources developed and released as a part of this competition, e.g. annotated data-sets and sentiment lexica [119]. In addition, there are many other linguistic resources that have previously been created and made publicly available for English SA, e.g. MPQA and Hu & Liu sentiment lexica [173, 97]. Benchmark studies are also of great importance in this context. For instance, a benchmark

analysis is conducted by Abbasi et al. [2] on a number of commercial and freely-available Twitter SA tools for English. However, the focus on English seems to shift to other languages: SemEval-2016 considers a sub-task for determining sentiment intensity of Arabic phrases taken from tweets. In sum, there is a need to handle different languages equally well, and hence, more research is needed to bridge this gap in Arabic SA.

While SA on longer and more structured text (e.g. web forums and reviews) has reached accuracy scores of up to 92.80% for English and 93.60% for Arabic [1], accuracy scores on Twitter messages are still far from that, with accuracy scores ranging between 65-71% on English tweets [2, 146, 145] and around 65.32% on Arabic tweets [8]. This is due to linguistic issues imposed by the social media text genre that result in difficulties in using existing NLP tools/techniques on low quality text, e.g. tweets [116]. Examples of these challenges are noisy, non-standard textual input and low-context language.

1.1.1 Thesis Goals

The main goals of this work are:

1. to empirically investigate and evaluate current SA techniques for Arabic (as an under-resourced language) and identify issues related to the Arabic language;

2. to determine the influence of novel feature-sets, data quantity and quality on the models' performance;

3. the provision and use of freely available data and tools.

1.2 Contributions

The main contributions of this thesis are as follows:

1. This thesis performs a systematic empirical evaluation of existing SA approaches. We identify existing SA approaches found to be successful on well-

resourced languages, and evaluate their effectiveness for Arabic tweets. This involves a principled comparison on benchmarking data and error analysis:

- We study the influence of extended features for SA in Arabic tweets. We show that a rich set of morphological features is amongst our best performing feature-sets. Other feature-sets that result in significant performance boost are semantic, affective-cues/social-signals and Twitter-specific features.

- We find that data quality is an important aspect in SA, with systems trained using manually annotated (gold-standard) data are able to attain promising performances. However, to cope with the evolving nature of Twitter, constantly obtaining training data manually is costly. We observe that turning to cheap alternatives (e.g methods that exploit emoticons) to automatically obtain large amounts of annotated data (data quantity) is viable, but influence the quality.

- We find that using a Machine Translation (MT)-based SA method, when no annotated data of sufficient quality and quantity is readily available, can eliminate the need for data annotation. The approach employs publicly accessible tools to translate Arabic tweets to English and utilises publicly available SA systems for English to label translated tweets.

2. Another outcome of this thesis is the provision of publicly released data-sets, each of which is annotated with a wide range of automatically extracted semantic, stylistic, Twitter-specific and morphological features:

   - A manually annotated gold-standard data-set of 9k tweets.

   - An automatically annotated emoticon-based data-set of 1.2M tweets.

   - An automatically annotated hashtag-based data-set of 130.2k tweets.

   - An automatically annotated data-set of 34,829 Arabic tweets. This data is labelled for sentiment by our best trained models.

   - A benchmark test-set of 3,538 diverse and fully-annotated tweets.

   - A manually annotated dialectal subjectivity lexicon of 489 items.

- A translated and manually filtered MPQA lexicon of 2,852 items.

- A manually annotated lexicon of social signals, i.e. prayers, regret, sigh, consent, dazzle and laughter.

3. We publicly release an SA tool for Arabic tweets that retrieves tweets from the live Twitter stream about a given query and automatically classifies them as positive, negative or neutral. An adapted version of this system is ranked top in SemEval-2016 Task 7 (Arabic Twitter subtask). This is the first time Arabic is considered in such an international competition for SA of social media data.

1.3 Thesis Outline

The thesis is structured as follows:

Chapter 2 (Background): This chapter first defines the concept of SA as a sub-task of text classification, and then lists the main potential applications and challenges of SA. It explains the main characteristics of the Arabic language and discusses the sources of difficulties when developing NLP applications for Arabic. Lastly, it identifies the most prominent approaches for SA that have shown to perform well on well-resourced languages.

Chapter 3 (Experimental Setup): This chapter describes the experimental set-up for the empirical work conducted throughout the following chapters. It describes the data-collection, text pre-processing and feature extraction. The chapter describes the process of annotating the corpus with a wide range of features, including: syntactic, semantic, stylistic, Twitter-specific and a rich set of morphological features. This is followed by an outline of machine learning schemes employed and evaluation metrics used.

Chapter 4 (Supervised Learning Approach): This chapter describes a supervised learning approach for SA on Arabic tweets. In addition, we study the individual contributions of various feature-sets. We also assess the performance of our feature-sets on a previously collected and manually annotated data-set of Arabic tweets.

Chapter 5 (Distant Supervision Approaches): This chapter investigates several methods for automatically creating training data, including emoticon-based, hashtag-based and lexicon-based distant supervision. The chapter presents an empirical evaluation of the performance of distant supervision for SA on Arabic tweets. The chapter also includes error analyses to identify major sources of errors with DS methods in Arabic. The chapter provides an analysis of learning rate in English vs. Arabic.

Chapter 6 (Machine Translation Approach): This chapter assesses a Machine Translation (MT)-based approach as a cheap and efficient alternative for SA, when no annotated data of sufficient quantity and quality is readily available. The chapter explores the scenario of translating Arabic tweets into English using publicly available MT tools and then employing off-the-shelf SA systems for English. Finally, we conduct an error analysis to understand the main sources of error in this approach.

Chapter 7 (Summary of SA Approaches): This chapter summarises the findings of the empirical investigations presented in chapters 4, 5 and 6. In addition, it presents the implementation of an SA system for Arabic tweets that exploits the best trained models.

Chapter 8 (Conclusion and Future work): This chapter concludes the work of this thesis, and summarises the main findings of each chapter. We also discuss possible future extensions for work presented in this thesis.

1.4 Thesis Publications

1. E. Refaee, and V. Rieser, (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. The 9th edition of the Language Resources and Evaluation Conference LREC' 2014. The European Language Resources Association. Reykjavik, Iceland 26-31 May 2014. (29 citations)[2]

2. E. Refaee, and V. Rieser, (2014a). Can we Read Emotions from a smiley face? Emoticon-based distant supervision for subjectivity and sentiment analysis of

[2]Google Scholar. Accessed on 21 July 2016.

Arabic Twitter feeds. In the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data. LREC' 2014. Reykjavik, Iceland 26-31 May 2014. (4 citations)

3. E. Refaee, and V. Rieser, (2014b). Subjectivity and Sentiment Analysis of Arabic Twitter feeds with limited resources. In Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools. LREC' 2014. Reykjavik, Iceland 26-31 May 2014. (15 citations)

4. E. Refaee, and V. Rieser, (2014c). Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds. In The Arabic Natural Language Processing Workshop ANLP co-located with EMNLP 2014 (Association for Computational Linguistics), Doha, Qatar 25-29 October 2014. (4 citations)

5. E. Refaee, and V. Rieser, (2015). Benchmarking Machine Translated Sentiment Analysis for Arabic Tweets. In the North American Chapter of the Association for Computational Linguistics - NAACL 2015 Student Research Workshop (SRW) (NAACL HLT 2015). Denver, Colorado, USA. 31 May - 5 June 2015.

Accepted for publication:

6. Refaee, E. and Rieser, V. (2016). iLab-Edinburgh at SemEval-2016 Task 7: A Hybrid Approach for Determining Sentiment Intensity of Arabic Twitter Phrases. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval'16, co-located with NAACL'16, San Diego, California, June 2016. **(Top system in the Arabic Twitter subtask)**

Chapter 2

Background

This chapter presents relevant background about the the key concepts related to work presented in this thesis. These include: sentiment analysis, social media mining, the Arabic language and prominent approaches previously employed for SA.

2.1 The Problem of Sentiment Analysis

SA is concerned with studying the sentiments, evaluations, attitudes and emotions conveyed in the form of written text [112, 164]. This definition is mirrored in the dictionary definitions of the word *sentiment* that include: a view or opinion that is held or expressed, an attitude toward something, a mental feeling, an emotion, an exhibition of feeling or sensibility, a thought influenced by or proceeding from feeling or emotion, e.g. a sentiment of pity. [1] As a research area, SA intersects with other disciplines, most importantly for this work are: natural language processing (NLP), machine learning (ML) and text mining (TM). NLP for social media is a relatively recent research area wherein the focus is on adapting traditional NLP approaches to the different text genre posted in social media [73]. Text mining is the process of discovering useful patterns, automatically or semi-automatically, in large quantities of data [175]. A sub-area of text mining is the text classification under which the sentiment analysis problem investigated in this thesis lies. Text classification is the task of assigning a text instance to one of predefined classes/categories [113]. SA

[1] Oxford Dictionaries. `http://www.oxforddictionaries.com/definition/english/sentiment`. Accessed on: 19 Oct 2015.

can be cast as a text classification problem wherein the task is to classify instances (e.g. tweets) based on the emotional orientation they convey to positive, negative or neutral.

Acquiring people's opinions, sentiments and evaluations has long been an active area of interest [112]. For example, organisations and businesses have always wanted to find out about how their products/services are being received by their customers [112]. For this purpose, conducting surveys and opinions polls has become a business itself for providing vital knowledge required for manipulating marketing strategies, managing political campaigns, and so on [125, 137]. In addition, individuals are interested to know what other people think about a certain product to help them in the decision making phase [75]. These opinions can be found as written text, e.g. a YouTube comment, a tweet, an SMS message or a product/movie review. The recent provision of such highly subjective text and the wide range of practical and industrial applications have motivated studies on SA, making it one of the most active research subjects in NLP in the past few years [112, 75].

2.1.1 Research on Sentiment Analysis

Research on people's opinions expressed in written text presents a highly challenging NLP problem that has been mainly approached in two major directions: sentiment analysis and emotion analysis [152]. For SA, the studies which have considered this direction have been mainly concerned with determining the sentiment orientation of a given text, i.e. positive, negative, or neutral (e.g. [1, 81, 37, 186, 160, 141]). The other research direction is emotion analysis in which work is concerned with identifying concrete emotion, i.e. joy, sadness, fear, anger, disgust and surprise (e.g. [135, 180]). This thesis focuses on SA. Although SA appears to be a less complex problem (binary or three-way classification) compared to emotion analysis, it is still challenging, particularly in noisy domains (e.g. Twitter) and less-resourced languages (e.g. Arabic). In addition, SA research has more potential in industry (e.g. assessing the success of a product), while emotion analysis seems to be more beneficial for social studies (e.g. assessing public mood/well-being).

2.1.1.1 Sentiment Analysis SubTasks

SA can be further broken down into the following subtasks: First, we can consider looking at sentiments conveyed as generic or topic-specific [152]. That is, the task can be defined as either determining the sentiment orientation of a given text in general, or with respect to a specific entity/aspect [145]. For instance, the SA system can be trained to decide if a given tweet is positive/negative despite the topic/entity about which the sentiment is conveyed, or if the tweet has a positive/negative sentiment towards some specific entity (e.g. a public figure) [34]. In this work, we study the overall sentiment polarity of a text instance [146]. Secondly, we can distinguish between studying sentiment from the author's perspective "who expresses the sentiment" (e.g. [135]) or from the reader's perspective "who reads the sentiment" (e.g. [166, 21]) [112]. The two subtasks can be mainly differentiated based on the way linguistic resources are annotated (e.g. training data). In this work, we are interested in exploring the conveyed sentiments from the author's perspective wherein each text instance is annotated with the intended emotion of its author in mind. Generally, most existing work of SA has considered the author's perspective [112], as the second subtask can be ultimately involved under the author's perspective, when readers turned their reactions/attitudes into a written form. In addition, Socher et al. [160] found that annotating text instances based on reader's perspective can lead to the majority of text considered neutral.

2.1.1.2 Sentiment Analysis Domains

The application of SA to various online domains has been investigated to replace methods traditionally used for collecting people's opinions (e.g. surveys and opinion polls) [137, 125]. Such domains include: newswire articles, newsgroups, reviews, forums and story sentences (e.g. [179, 136, 1, 126, 174, 165]). These domains can be characterised as being relatively formal and longer pieces of text, i.e. consisting of several sentences and/or paragraphs. The application of SA moves towards less formal domains. Examples of these domains are SMS messages, tweets and Facebook posts, which represent an informal text genre that introduces new challenges for

NLP, such as containing abbreviations, ungrammatical and incomplete sentences and spelling errors [186, 123].

2.2 Mining Social Media for Sentiments

Due to the free style of communication and the ease of accessibility, social media platforms have attracted a wide spectrum of internet users. People tend to use social media primarily either as a source of information or to share their thoughts and opinions [61, 120]. Therefore, social media content leverages the perspectives of millions of people that can be readily harvested and exploited for SA research and applications [73].

2.2.1 Challenges of Social Media Data

The difficulty associated with processing social media is mainly due to the fact that many of the existing NLP tools and resources have been developed for and evaluated against a formal/edited form of text (e.g. newswire) [146, 123, 54]. A considerable body of literature has recently identified a number of possible sources of difficulties associated with social media [136, 116, 117, 135, 120, 13, 54, 115, 61, 82, 65, 73]. Such challenges involve:

- being informal, i.e. written in non-lexicalised form, such as: the non-standard use of punctuation, lengthening (e.g. happppy), abbreviations, creative spelling, misspellings, slang and swear words.
- being non-grammatical, i.e. not as thoughtfully composed or edited as in traditional media sources, and, hence, sentences can be poorly structured and have grammatical errors.
- may also convey sarcasm, irony, mixed and/or unclear polarity content that pose a challenging task for ML techniques to recognise.

In this context, *"bad language"* used within social media is defined by Eisenstein [61] as "the text that is associated with noise with respect to its non-standard spelling,

and syntax". He highlights social media's role in provoking the users' desire for self-presentation that results in the diversity encountered in social media language. Eisenstein [61] suggests that normalising text (i.e. mapping to known words) to fit tools/systems designed for a more standard form of text might reduce the meaning of the text. For example, the direct mapping from *bro* to *brother* might eliminate the negative impression that *bro* can create. In accordance with this view, we present a set of text pre-processing procedures that we have employed for the experimental work described in this thesis (section 3.2).

The use of sarcasm and mixed sentiments is one of the most difficult problems for SA, as it might influence the sentiment orientation of text [136, 112]. Tweets with mixed sentiments provide a common source of SA errors [2, 123, 146]. Maynard et al. [116] note that tweets tend to contain extensive use of irony and sarcasm. Sarcastic tweets are reported to represent up to 13.5% in a data-set of Arabic tweets [120]. Other studies chose to simply exclude sarcastic instances (e.g. [99]). We have manually examined samples of our Twitter data-sets for sarcasm and note a tendency to convey a negative sentiment in a seemingly positive context, which is also noted by Maynard et al. [116] on English tweets. Table 2.1 (page 14) shows examples from our data-set. While examples 1 and 2 seem to have unclear sentiment orientation, examples 3,4 and 5 show cases of potentially sarcastic views being expressed, and examples 6,7 and 8 indicate mixed emotions. In addition, and in contrast to topic classification, sentiment can be expressed in an indirect manner (e.g. by avoiding the explicit use of negative words), making them harder to be identified [128] (as in example 9). Another issue we observe is the use of positive words in a negative context or vice versa (as in example 10), an issue which is also noted by Mourad and Darwish [120].

The aforementioned characteristics demonstrate micro-blog text genre's substantial difference from traditional edited text and raise the need for domain adaption, which involves developing linguistic resources that particularly target this text genre [61, 73]. Such resources include Twitter-tuned NLP tools, e.g. NER [144, 52], Twitter sentiment data-sets and sentiment classifiers [81, 58]. One of the main goals of

(1)	المَساوَاه في قمع الحريَات الشخصية عدل *Equality in suppressing personal freedom is justice*
(2)	احيَانَا فهمنَا الَّامور بطريقه خطَا يكون هو الصح *Sometimes, the wrong understanding of things leads to the right thing.*
(3)	مصر دلوقتي بقت عَامله زي الفِيلم الَاجنبي الغير مترجم، الكل بيتفرج ويترجم علي مزَاجه *Egypt now is more like a foreign film without subtitles, so everybody watches and puts their own translation.*
(4)	الكويت مركز انسَاني!!! هم يستهبلون اكيد *Kuwait is a centre for humanity!!! they must be kidding.*
(5)	مبروك خسرت كرَامتك *Congratulations! you have lost your dignity.*
(6)	لست مع الَاخوَان سيَاسيَا، ولكنَى معهم انسَانيَا *I disagree with Muslim-Brotherhood politically, but support them humanly.*
(7)	السنه و الشيعه كل طرف يحمل صوره نمطيه عن الَاخر فيهَا الكثير من الزيف و الحق *Sunna and Shiah (major sects in Islam), each holds a stereotype/received idea about one another with lots of falsity and truth.*
(8)	القذَافِي كَان اهبل، بس محترم *Gaddafi was insane, but respectable.*
(9)	بعض النَاس مثل الضفدع، حتي لو رفعته علي كرسي الذهب، سيقفز الي المستنقع مجدّدًا *Some people are just like frogs, even if you put them on a golden chair, they will jump back to the swamp.*
(10)	العرض رَاءيع بطريقه مرعبه *The show is gorgeous in a (terrifying) way.*

Table 2.1: Example tweets with undeterminable, mixed sentiment indicators, or a potential use of sarcasm/irony.

this work is to develop and provide freely available data/tools/resources for SA in Arabic.

2.2.2 Sentiment Analysis in Twitter

Twitter is among the best recognised micro-blogging platforms around the world [44] (see figure 2.1) and has attracted a considerable amount of research in various NLP applications, including SA (e.g. [81, 37, 127, 186, 135, 77, 120, 123, 146]). SA for Twitter is not a trivial task due to the complexity and variability of sentiment indicator(s) that a single tweet can contain [37].

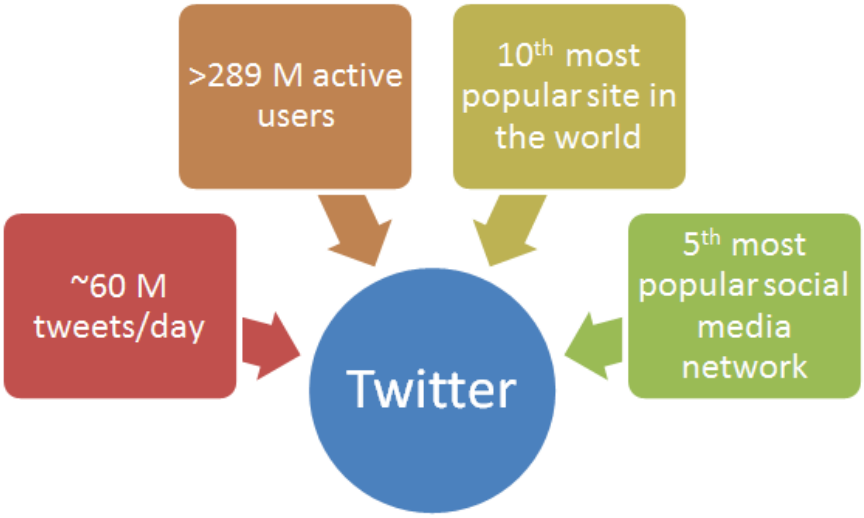


Figure 2.1: Twitter Stats [161].

As a micro-blogging service, Twitter has an inherent number of challenges typically associated with the social media text genre, as discussed in section 2.2.1. In addition, its posts are usually used for sharing information and opinions are characterised by their short length (limited to 140 characters). As such, the possibilities of encountering/capturing sentiment-bearing words, which are crucial for most (if not all) SA systems, decrease notably. For instance, although Taboada et al. [165] use a sentiment lexicon with nearly 5k entries, the authors notice a significant increase in the number of empty text,[2] from 0.2% in movie reviews to 21.7% in blog posts. In

[2]Empty text is a text instance containing no sentiment-bearing words matching any of the words in the dictionaries used [165].

addition, the issue of limited length can pose further issues, such as the heavy use of non-standard/creative/improvised spellings (e.g. *ugh*, *ew* or *sux* instead of disgusting), abbreviations and limited contextual information that can block the necessary clues that would otherwise help deciding on the overall sentiment orientation of a tweet [115, 50, 116, 73]. The non-standard spelling can be seen as an emerging means for conveying layers of meaning in a way that copes with the requirements of social media interactions [61].

A relevant issue is the dynamic/time-evolving nature of the Twitter stream as people can discuss a large number of different topics, which also has an impact on the rate of lexical variation [37, 109, 82]. Eisenstein [61] studies several Twitter data-sets and observes that the proportion of the out-of-vocabulary (OOV) bi-grams increases over time. Dodds et al. [58] conducted a large-scale study on 24 corpora in 10 languages, including Arabic, from several sources, including Twitter, New York Times and movie subtitles, and observe that the Twitter corpus is the most variable one. A possible solution to overcome the high rate of lexical variation is by exploiting large amount of tweets (chapter 5) that can be freely crawled and capture as many of the linguistic characteristics distinguishing the tweets' text genre as possible.

Another problem in the Twitter stream is redundancy. Examples of redundant content are re-tweets and repeated tweets (mostly advertised content). Sharifi et al. [158] observe that the highly redundant content in the Twitter stream is a major challenge for data mining tasks. To account for this issue, we follow a number of cleaning-up steps during the data collection phase (described in chapter 3).

Why Twitter?

Tweets are rich in linguistic variation making Twitter data-sets essential elements of web corpora with many research and application purposes. In this work, we focus on studying sentiments expressed in Twitter as it incorporates several advantages:

- Abundance of freely available, up-to-the-moment, and easy to obtain interactions as user-generated content, facilitating the build of large corpora.

- Micro-blogging platforms like Twitter are identified as one of the most popular categories of social media that internet users prefer over other communication

medium, such as mailing lists [127, 73].

- The enormous variety of its users' social, cultural and political backgrounds who tend to use Twitter not only for social networking, but also as a source of information [92, 33].

- A highly effective means for promoting (e.g. Twitter's role in the Scottish referendum campaign).[3]

- The influence/importance of social networks like Twitter in public life has grown notably. For example, a story published by BBC News revealed that:

  "*A Tory council candidate has resigned from the party with immediate effect after posting anti-Islamic and homophobic comments on Twitter*".[4]

2.3 The Arabic Language and its Presence in Social Media

Arabic is the language of an aggregate population of over 422 million people, the first language of the 22 member countries of the Arabic League and the official language in three others [169]. Arabic can be classified with respect to its morphology, syntax, and lexical combinations into three major levels: Classic Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) [83]. CA can be found in religious text, while MSA is the official language of education and media due to its standardisation [184]. DAs or the spoken varieties of Arabic are primarily used in informal daily communication as "the true native language forms" [83]. Habash et al. [88] define DA as "the day-to-day native vernaculars/dialects spoken in the Arab World". Habash [83] identifies five major groups of dialects in Arabic: Egyptian (includes Libyan and Sudanese), Levantine (includes Lebanese, Syrian, Palestinian and Jordanian), Gulf (includes Saudi Arabia, Qatar, Kuwait and UAE), Iraqi and Moroccan.

[3] `http://www.independent.co.uk/news/uk/politics/scottish-independence-twitter-weighs-in-with-reasons-to-vote-yes-9212749.html`. Accessed on: 21 Oct 2015.

[4] `http://www.bbc.co.uk/news/uk-politics-27272907`. Accessed on: 4 May 2014.

As social media has spread, DAs have found their way for the first time into written form and become digitally stored [101]. Whilst highly individual-driven, Arabic on social networks uses informal styles that are a mixture of local dialects, such as Egyptian Arabic and Gulf Arabic, side by side with MSA [71, 17, 88, 101].[5] Therefore, online communications (e.g. micro-blogs) represent a rich resource of the variable forms of the Arabic language, i.e. MSA, DAs or a combination of both, that can be exploited in creating data-sets for computational linguistic.

2.3.1 Why an Arabic Corpus of Social Media Content?

Given the recent political unrest in the Middle East (2011), there has been an increasing interest in harvesting information written in Arabic language from live online platforms, such as Twitter [44]. Social media has played a vital role in sparking the recent social movements in that central part of the world. As such, the Arab Spring and other ongoing conflicts and political movements have yielded heavy use of Twitter, and other micro-blogging platforms, to convey complex emotions reflecting personal stances towards such circumstances [90, 25, 90]. A recent study by Buettner and Buettner [44] reveals that Twitter has played an effective role in sociopolitical revolutions (e.g. the Arab spring) by allowing people to share their feeling of discontent, which has subsequently led into triggering political revolutions.

Another reason is that Arabic is one of the fastest-growing languages on the web [108]. With regard to Twitter, Arabic represents nearly 6% of the Twitter stream with the total number of active Twitter users in the Arab world reached more than 5M users, with an estimated average number of 17M tweets per day [118, 24]. While there is a growing interest within the NLP community to build Arabic corpora by harvesting the web, (e.g. [17, 4, 184, 184]), these resources have not been publicly released yet (section 2.3.2). I therefore built newly collected data-sets of Arabic tweets annotated for SA (chapter 3). Al-Twairesh et al. [19] state that the availability of annotated corpora, which is a necessity for SA systems, is still scarce for Arabic. The authors point out that the first release of my corpus [138]

[5] Arabic represents an example for the interesting phenomena of ***diglossia*** where two varieties/forms of the same language live side-by-side [76, 188, 71].

forms part of the recent efforts to tackle the issue of the lack of Arabic data-sets annotated for SA and targeting domains other than news and movie reviews.

2.3.2 Current Efforts to Develop NLP Tools and Resources for Arabic and its Dialects

The formal variety of the language, namely MSA, has been the subject of considerable efforts in developing NLP tools spanning various aspects, such as: tokenisation, POS, stemming and machine translation. Habash [83] reviews a set of the most popular tools and systems that has shown to be of great value for Arabic NLP. In contrast, NLP research on Dialectal Arabic has only recently flourished to cope with the increasing prevalence of DAs on the web. DAs differ significantly among themselves and from MSA [185], and, hence, each variety can be treated as an independent language [184]. One of the biggest challenges facing DA research is the lack of annotated resources required for building robust NLP tools and applications [85]. For instance, the Linguistic Data Consortium (LDC) catalogue lists 91 linguistic resources for MSA compared to 15 for Egyptian, 5 for Gulf, 3 for Moroccan and 13 for Levantine.[6] Responses by the research community to address this issue take two major directions: building new linguistic resources and tools for DA; or extending existing state-of-the-art tools to cover one or more of the local dialects. In this section, we review some of the most prominent efforts in this area.

A) **Corpora Building:**

Exploring the existing work on building Arabic corpora revealed YADAC (Yet Another Dialectal Arabic Corpus) [17], a multi-genre dialectal Arabic corpus. YADAC includes web data from micro-blogs, blogs, forums and online market services [17]. However, it is designed for Egyptian Arabic only and has not been made public yet.

Abdul-Mageed and Diab [4] presents AWATIF, a multi-genre corpus that is manually labelled for SA.[7] The corpus is a collection of newswire stories, Wikipedia

[6] `https://catalog.ldc.upenn.edu/search`. Accessed on 03 Nov 2015.
[7] AWATIF is a transcription for an Arabic word meaning sentiment.

talks (political, religious), and forum conversations. However, the corpus does not include any social media posts (e.g. tweets) and the authors excluded non-MSA from their data-sets.

A recent publication by Ibrahim et al. [98] presents MIKA, which is an Arabic corpus that comprises different text genres, including tweets, reviews and comments. Although the tweet portion of corpus is relatively small (only 1k tweets), it can be utilised (e.g. as a validation/development set), especially that it is manually annotated for SA by three native speakers of Arabic. A limitation is that the corpus targets MSA and Egyptian dialect only. In addition, the authors excluded instances with mixed or sarcastic views. This is a shortcoming, as such views are important perspectives to be considered with SA problems.

B) **Sentiment Lexica:**

Abdul-Mageed et al. [7] present ArabSenti, an Arabic subjectivity lexicon with around 3600 words. In addition to being publicly shared, the lexicon is manually annotated for sentiment by two native speakers of Arabic. As such, we believe ArabSenti is a valuable resource that we exploit in our experiments (i.e. for extracting semantic features).

In recent work, Abdul-Mageed et al. [5] describe their ongoing efforts to build a multi-genre sentiment lexicon for Arabic. However, the lexicon covers only two dialects besides MSA - Egyptian and Levantine Arabic - and has not yet become publicly available. In addition, being a large-scale (more than 200k) and automatically generated lexicon, it might suffer from some degree of noise and "too much coverage" as a result of assigning sentiment orientation for words which, in fact, are neutral [165].

Another recent attempt by Eskander and Rambow [68] presents SLSA, a Sentiment Lexicon for Standard Arabic. SLSA is a large-scale sentiment lexicon for Arabic wherein each entry is associated with a Sentiment Intensity (SI) score indicating its strength of evaluative intensity. The scores are assigned using a linking algorithm that links the English gloss of each entry to a synset from

a large-scale sentiment lexicon for English that is associated with an SI score, namely SentiWordNet [69]. SLSA has advantages of being made publicly available, and comprises nearly 35k lemma. However, SLSA covers only MSA.

C) **Orthography Standardisation:**

To partially address the issue of lack of standard orthography of DAs, Habash et al. [85] propose a conventional orthography for dialectal Arabic, namely CODA. The authors suggest that CODA would be able to successfully solve the spelling inconsistency and sparseness problems by mapping various spelling-variants onto a single orthographic form, leading to a more robust language model. However, the current version, CODAfy, was built with only Egyptian dialect in mind [86] and has not become publicly available yet.

D) **Morphological Analysis:**

Tokenisations, POS, morphological analysis and disambiguation are essential to support higher order NLP applications like text classification [83]. Consequently, a considerable effort has been made to develop tools and resources tackling these aspects for Arabic, with a particular focus on MSA [87, 124, 148]. In particular, researchers at Columbia University have developed a state-of-the-art automatic morphological analyser for Arabic, namely MADA (Morphological Analysis and Disambiguation of Arabic). It is also called MADA+TOKAN, as it incorporates a morphological tokeniser. This system derives a linguistic interpretation/analysis of each word for a given Arabic text [87, 124]. As the importance of studying DAs increased, researchers have investigated the possibilities for extending such a valuable tool for morphological analysis in MSA to include DAs [88, 56, 131]. Therefore, Habash et al. [88] propose an extension for MADA to cover Egyptian Arabic. For the purpose of extending MADA, they have utilised recently developed morphological analyser for Egyptian Arabic and a large Egyptian corpus annotated morphologically [86]. In a later study by Pasha et al. [131], the authors propose another enhancement for MADA by merging it with a previously built and commonly used system called AMIRA [57]. They state that merging

the best features of MADA and AMIRA will allow the expansion of the capabilities of the newly merged system, namely MADAMIRA, and allow a significantly faster system. In this work, we employ MADAMIRA (V.1.0) to extract a rich set of morphological features (chapter 3), accounting for the morphologically-rich nature of Arabic. Although MADAMIRA (V.1.0), with the extension of the Egyptian dialect, has been officially released, obtaining the Egyptian components requires having an LDC licence. One of the main goals of this work is to utilise publicly available tools. Therefore, we use the freely available version of MADAMIRA, which supports MSA only, even though we anticipate this will introduce some noise with the extracted features and will possibly have a negative impact on accuracy. This is an issue that we empirically investigate in this work (chapter 4).

E) **Language Variety Identification:**

Zaidan et al. [183, 184] present their effort to build a new Arabic resource, the Arabic Commentary Data-set (AOC), with dialect annotations at sentence-level by crowdsourcing. The data-set is harvested from reader commentary on online newspapers and manually annotated for dialect to train classifiers for automatic dialect identification.

In subsequent work by Elfardy et al. [67] and Elfardy and Diab [66], AOC is used in a supervised-based system for dialect identification at sentence-level. The resultant dialect-identification system is called AIDA. The authors reported an accuracy score of 85.5% for MSA vs. DA task.

The work of Zaidan et al. [183, 184] is further extended by Cotterell and Callison-Burch [50] who built a human annotated corpus with data obtained from newspaper commentary and Twitter. The corpus is freely available and has the advantages of involving Twitter data and being manually classified into five Arabic dialects using Amazon MTurk. Training NB and SVM classifiers with n-gram features and 10-fold CV setting, the authors reported accuracy scores of up to 87% on Gulf Arabic and 84% on Egyptian and Levantine Arabic.

In this work, we use AIDA, as a publicly available tool, to extract the language-class of a tweet, i.e. MSA or DA. This is because AIDA is a ready-to-use system that comes with its built models. In addition, AIDA can extract the tweet's degree of dialectness. We exploit these two features (i.e. language-class and degree of dialecness) under the language-style feature-set (see section 3.3). The data-set of Cotterell and Callison-Burch [50] will require training new models and is not able to obtain degree of dialecness. Nevertheless, this data-set has the potential to be used in future investigations of per-dialect SA, discovering issues like whether a certain dialect is more/less difficult for SA. In this work, we consistently study multi-dialect data-sets.

F) **Machine Translation:**

The MSA variant is handled well with existing MT systems (e.g. Google Translate) [182]. However, other Arabic variants (e.g. Levantine Arabic) are still not as well handled [84]. That is, attempting to translate a DA text using an MT system trained on MSA data, which is the case with most (if not all) of the existing publicly available MT systems (e.g. Google Translate) [73], is likely to result in failing to translate parts of text or translation errors [182]. A main reason for this issue is the lack of resources necessary to perform MT on DAs, like parallel corpora (e.g. dialect-English) [185, 56].

MT on DAs has been addressed in literature in two major directions 1) mapping DA to MSA and then applying an existing MT system [154, 155]; 2) considering each dialect as an independent language and creating new linguistic resources (e.g. parallel corpora) accordingly [185]. For the first approach, Salloum and Habash [154, 155] present a rule-based system called Elissa (v1.0) that uses several components like language models and dictionaries to produce MSA alternative normalisations of given dialectal sentences. The current version of Elissa targets Levantine and Egyptian Arabic and has not become publicly available yet. In the second approach, Zbib et al. [185] considered crowdsourcing to build two parallel corpora: Levantine-English and Egyptian-English. Data was sampled from a corpus of Arabic web text. The resultant translated corpus includes

1.1M Levantine words and 380k Egyptian words. Such a corpus is promising for building MT systems for DA, but accessing the corpus requires an LDC subscription (i.e. not freely available). By the time we conducted the empirical investigations in this work, there was no freely-available MT system for DAs. Thus, in chapter 6 we investigate the impact of employing publicly available MT tools (i.e. Google and Bing) to translate Arabic tweets into English and assess the performance of SA systems on translated tweets.

2.3.3 What are the challenges of SA in Arabic?

A major challenge with Arabic NLP is the rich morphology of Arabic language [83, 8, 13, 3]. As a morphologically-complex language, Arabic includes rich inflectional morphology where a significant amount of information is expressed in units of a word affecting its overall behaviour syntactically and semantically [83, 8, 88]. That is, Arabic has eight obligatory inflectional features for every word [83]. These features are: aspect, mood, person, voice, case, state, gender and number (further discussion in chapter 3). The variation across the inflectional features of a word yields a significant increase in the number of possible word forms [120, 19], which increases data sparsity, and, hence, makes word-based data-driven approaches to SA more challenging [3]. As one lemma can be associated with thousands of surface forms, a possible solution is in reducing word forms into a more compressed form (e.g. stem or lexeme) [8]. Despite its potential for vocabulary reduction, the solution of using compressed forms might imply some information loss (i.e. collapsing/destroying many sentiment distinctions). In chapter 3, we describe our preliminary experiments for identifying the potentially most useful word-token-based form for the task of SA (e.g. normalised surface form, lexeme, stem, etc.). As a compensation for the possible loss of information when employing a reduced form of words, we automatically extract the full set of inflectional features for each word and exploit them as discriminative features for training the sentiment classifiers (i.e. the morphological features extracted using MADAMIRA, see page 62).

Additionally, SA on Arabic social media posts (e.g. tweets) is a challenging task

due to two reasons: 1) the limited availability of resources and 2) the additional need to tackle the use DAs. The limited availability of linguistic resources (i.e. annotated data-sets and sentiment lexica) is a major challenge facing SA on Arabic [13, 7, 120, 64]. Despite the recent interesting efforts to address this issue (e.g. [7, 4, 120, 5], see section 2.3.2), Arabic remains an under-resourced language with respect to SA resources compared to English. Al-Twairesh et al. [19] state that current SA resources suffer from low accuracy and are tailored for specific dialects. By the time of this thesis, there was no publicly available corpus of Arabic tweets annotated for SA.

DAs create additional challenges for researchers working on NLP: as they are mainly spoken dialects, they show significant variation from MSA and lack standardisation [83, 184]. To illustrate, DAs have no standard orthographies [88, 101], which implies that the problems of free-writing style and improvised spelling (section 2.2.1) are even more pronounced with Arabic SM posts [185, 67]. This issue has triggered interesting recent efforts to standardise DA orthography (section 2.3.2). However, such tools have not become publicly available. As such, the empirical work described in the following chapters considers analysing Arabic tweets without applying spelling adjustment for the dialectal words, so that we explore the robustness of the trained SA classifiers in spite of such noise.

Another issue with Arabic is the conventional omission of short vowels that results in confusion among MSA and DAs. For instance, dialectal and MSA words might have the same spelling with different meanings and phones, namely homographs [184]. Zbib et al. [185] and Elfardy et al. [66] find this challenge among the most problematic cases for NLP tasks on DAs, e.g. machine-translation and dialect identification (see examples in table 2.2). To account for this, we employ a rich set of POS tags, automatically extracted and utilised as features, to learn SA classifiers using a state-of-the-art morphological analyser for Arabic. In addition, we employ a binary feature that identifies whether a given tweet is MSA or DA.

Another problem that can be drived from informality associated with social media text genre (see section 2.2.1) is when bi/multi-lingual users post text on

(1)	هم	an MSA noun meaning *worry/terrible*, or an adv meaning *also* in Gulf Arabic.
(2)	مرتبك	an MSA adj meaning *confused* or an Egyptian noun meaning *your salary*.

Table 2.2: Examples of homograph words in MSA and DAs.

social networks and use a mixture of languages [64, 105], as in the example (1) in table 2.3 taken from our corpus. Furthermore, users may tend to introduce their own abbreviations either to cope with the requirements of social media or as a part of showing their identity/style [61], as in example 2 in table 2.3. These issues accumulate to: firstly, the problem of expanding the space of lexical variety that increases sparsity, especially with the word-token-based features; and secondly, the degree of noise encountered with the extracted features (e.g. morphological features).

(1)	فَالَنتَاين، كيوت، ثَانكس	*Valentine's day*, *cute* and *thanks* in English, spelled using the Arabic alphabet respectively.
(2)	برب	*I'll be right back*

Table 2.3: Examples of transcribed English words appear in Arabic alphabet.

In the investigations presented in this work, we do not perform any normalisations to the aforementioned challenges (i.e. to correct misspellings or map/translate dialectal words into MSA). Instead, we apply pre-processing/preparation techniques to tweets, e.g. by reducing expressive lengthening (page 56). The purpose is to assess the ability of SA classifiers to deal with noisy live data from the Twitter stream.

2.4 Sentiment Analysis: Prominent Approaches

The literature on SA reveals two major approaches for SA [128]: The first one is the *lexicon-based* approach that exploits word dictionaries, with words being associated with their prior sentiment polarity [11, 10]. This approach may also incorporate intensifiers (e.g. very) and shifters (e.g. negators) to calculate an overall polarity score for a given text. The second major approach is *machine-learning-based* wherein annotated examples/instances are used to learn a model to automatically identify the sentiment orientation of a given text. In this section, we review example previous work that adopts either these approaches or a combination of them to carry out SA with a particular focus on the domain of social media. The purpose of this section is to provide an overview and point out general shortcomings of existing work on SA for Arabic tweets. In addition to the related work presented in this chapter, we will also review relevant approaches in individual chapters.

2.4.1 Lexicon-based Approach

The lexicon-based approach for text classification problems involves calculating an aggregated score for a document based on the presence of words from dictionaries with values – manually or automatically - assigned as positive (e.g. brilliant) or negative (e.g. horrible). The resultant score is then used to automatically assign the text instance with a sentiment label (i.e. positive or negative).

Two main methods have been used to handle negations in lexicon-based methods: 1) *negation switch* and 2) *negation shift*. Negation switch is the most popular method in which the polarity of the following sentiment-bearing word is simply reversed (e.g. from positive to negative). Negation shift is a more sophisticated method in which the presence of a negator will result in subtracting a fixed value from the following sentiment-bearing word (i.e. making a word less positive rather than totally flipping polarity to negative). Overall, Taboada et al. [165] experimented using both negation settings and the results reported show a small difference between the two methods, with negation shift being slightly better.

Read and Carroll [137] present an investigation of a lexicon-based method for SA on a data-set of English movie reviews. The authors explore the performance of three word similarity techniques to automatically build different sentiment lexica. The authors reported the best F-score at 0.687 for positive vs. negative using word-lemma as features.

Taboada et al. [165] present a lexicon-based system for SA in English. The proposed system incorporates semantic orientation of individual words and contextual shifters (e.g. negators). The authors highlight the fact that the reliability and quality of dictionaries of sentiment-bearing words used are crucial for building a robust SA system that uses a lexicon-based approach. Unlike Read and Carroll [137], who used auto-generated sentiment lexica, Taboada et al. [165] argue that the noise introduced by auto-generated sentiment lexica makes dictionaries employed for lexicon-based approaches less reliable. Therefore, they built their own manually annotated lexicon that includes nearly 5k words, with each word being assigned with a hand-ranked sentiment orientation value (positive or negative). They excluded neutral words from the dictionaries. To improve coverage, they lemmatised words in the lexicon used for SA, like Read and Carroll [137]. The authors reported an average accuracy score of 78.74% on reviews and 75.31% on blog posts for positive vs. negative. They conclude that larger and auto-generated dictionaries are not necessarily better, mainly because they introduce more noise. This is due to the fact that many of the words in these dictionaries will be assigned scores, even though they are not sentiment-bearing words. The issue of data quality vs. quantity is an aspect that we assess its impact in this work for Arabic SA.

As for Arabic, Abdulla et al. [9] report on their investigations to perform binary classification (positive vs. negative) on Arabic tweets using a lexicon-based approach. They used a manually annotated lexicon to extract sentiment-bearing words from tweets and assign them with a polarity score (-1 or +1). The aggregated score is then used to assign a sentiment label to each given tweet. Comparing the auto-generated labels to manually assigned ones, the authors report an accuracy of 59.6% and an F score of 0.616. The data-set used is relatively small (2k tweets) and

only includes MSA and Jordanian (part of Levantine Arabic) tweets.

In a recent study, Wang et al. [171] developed a system for SA on Arabic tweets that employs a lexicon-based method. The lexicon used was created by translating an English lexicon and manually filtering irrelevant entries (we follow a similar approach in our work, see page 50). The proposed system targets Egyptian and Saudi dialects; hence, they expand the lexicon by manually adding the equivalent of each entry from both dialects. Using a set of selected keywords, the authors collected tweets about three topics they were interested in: Egyptian government, telecommunication and employment. The system was tested on a small set of 1200 manually annotated tweets and the average observed F-score is 0.801. Although results are generally better than those reported by Abdulla et al. [9] on Arabic tweets at 0.616 F-score, the superiority can be partially attributed to the fact that the collected data only took three topics into account, which is likely to reduce sparsity and improve performance [137]. In this context, Abbasi et al. [2] run a cross-topic SA experiment on tweets and observe that results varied across test-sets (i.e. topics). That is, better results are expected, even on highly noisy and short pieces of text like tweets, when tuning a system to be a topic and/or dialect specific [171, 167, 22]. In our work, we explore the robustness and effectiveness of sentiment classifiers that we train on general-purpose and multi-dialectal data. Our aim is to produce a system that tackle any random sample taken from the Twitter stream.

2.4.2 Machine Learning Approaches

Machine learning involves using a machine learning scheme to learn a statistical classifier. This is accomplished by presenting the classifier with a set of labelled examples from which it is expected to learn to classify unseen examples. The classifier's ability to classify previously unseen instances is called *generalisation* [175]. With the actual outcomes/classes being pre-defined, this method of learning is known as *supervised learning*.

Supervised learning (SL) techniques can perform the task of sentiment classification effectively within various domains. These involve: newswire articles [179],

movie/product reviews [129, 47, 1], web forums [1], question-answering opinion corpora [174] and social media (e.g. Facebook and Twitter) [100, 120].

As for SA in Twitter, Nakov at al. [123] created a data-set of nearly 10k manually annotated English tweets. The annotation was done at two levels: phrase-level and tweet-level. The data-set was used in a shared task in SemEval-2013 with two main sub-tasks.[8] Sub-task A is concerned about determining the sentiment orientation at phrase level, while sub-task B is concerned about identifying the overall polarity of a given tweet, i.e. positive, negative or neutral. In this work, we focus on reproducing sub-task B for Arabic tweets. Nakov at al. [123] report sub-task B to be more difficult than sub-task A, attributing this difficulty to the presence of instances with mixed emotions when considering the entire tweet. In the later editions of the shared task [146, 145], further data-sets were released. Overall, they report that almost all systems used SL. The best results were reported to be achieved using popular machine learning schemes that have shown to be successful on text classification tasks, such as Support Vector Machines (SVM) and Naïve Bayes (NB). The reported F-scores on sub-task B in SemEval-2015 are between 0.648 and 0.248.

With regard to Arabic, Farra et al. [72] experimented on a set of nearly 2k Arabic movie reviews that were manually labelled. By training an SVM classifier and utilising a set of semantic (e.g. presence of positive/negative words) and stylistic (e.g. presence of special characters) feature-sets, the authors report accuracy scores of up to 84% on the binary classification (positive vs. negative). Similarly, Abdul-Mageed et al. [7] presented their experiments on a manually annotated MSA newswires. The corpus is annotated at the sentence-level of more than 2k sentences. By training an SVM classifier, the authors report an F-score up to 0.955 for the binary classification (positive vs. negative) and 0.715 for polar vs. neutral when combining a set of syntactic (word-stems), morphological and semantic features.

Although results on reviews and newswires appear promising, more investigations are needed to explore the effectiveness of standard approaches (e.g. SL) and features (e.g. syntactic and semantic) on different text genres (e.g. social media). For this

[8] SemEval's SA task on tweets is the most popular SA shared task to date with more than 40 teams from allover the world participating in each of the previous three editions of the task [145].

purpose, Abdul-Mageed et al. [8, 6] present a system for SA on Arabic social media content incorporating text genres like tweets and web forums. They built a manually annotated data-set of 3k tweets and train an SVM on set of word-based n-grams and semantic features. The authors reported the best results on the Twitter data-set at 65.87% accuracy and 0.618 F-score for positive vs. negative on a held-out test-set that is a split of the original training data-set.

Another work by Mourad and Darwish [120] conduct a set of investigations on a collection of 2k manually annotated Arabic tweets. The authors utilise a set of syntactic (word-stem), semantic, POS, stylistics (e.g. presence of punctuations) and Twitter-specific features (e.g. presence of hashtags). Training NB and SVM classifiers, they report scores of 71.9% accuracy and 70.35% F-score for positive vs. negative. Although results are generally better than those reported by Abdul-Mageed et al. [8] on a data-set of Arabic tweets, it is worth mentioning that the experimental setup of Mourad and Darwish [120] used a cross-validation configuration, whereas Abdul-Mageed et al. [8] used a subset of data to evaluate the SA models against, meaning that there is still a need to test the trained models on a test-set that is collected at a later point in time to explore the performance of SA models for a dynamic medium (e.g. the Twitter stream).

Conclusion: As it can be seen, the results are generally far below those reported on Arabic movie reviews (84%) and newswires (95%), suggesting that further investigations are required on such a noisy text genre. In this work, we experiment with expanded, more variant feature-sets and larger training data. In addition we empirically evaluate and compare multiple existing approaches for SA.

SL approaches face the challenge of limited availability of labelled data for training and evaluation [82, 64]. With a data-streaming medium like Twitter, manual sentiment labels are not only expensive to obtain, but also become unpractical with millions of tweets generated everyday. This makes keeping models up-to-date with manually annotated data a hard task [186]. A possible remedy is by exploiting techniques for automatically obtaining labelled data for training. These "techniques one can use to try to obtain the benefits of considerably larger training corpora without

incurring significant additional costs of manual annotation" [31]. A range of hybrid methodologies, also known as *distant-supervision* approaches, have been proposed in the literature, bringing together machine-learning-based and lexicon-based approaches, among others.

2.4.3 Distant-Supervision Approaches

Researchers have investigated ways for obtaining sentiment annotation in a timely manner along with no, or at bare minimum, level of human intervention, such as Distant supervision (DS) approaches. DS exploits existing features, e.g. emotions or sentiment-bearing words, to automatically annotate training examples. Although DS can yield noisy labels, it can provide a cheap and effective way to directly access the author's emotional state. This section presents two existing hybrid approaches:

1) The combination of conventional markers and machine-learning approaches,

2) The combination of lexicon-based and machine-learning approaches.

2.4.3.1 Conventional Markers + Machine Learning

This approach leverages existing conventional markers, such as emoticons and sentiment bearing hashtags, as noisy labels to build training data automatically, as first proposed by Read [136]. Emoticons are visual cues that are assumed to be used by users to mark up their own emotional state expressed in the accompanying text [135], and presumably used as author-provided sentiment indicators. Hashtags represent another conventional marker of Twitter that is used to filter tweets according to a keyword or topic (i.e. #Syria). Emoticons and sentiment-baering hashtags are merely used to collect and build the training data. To avoid biasing the data and to force the classifiers to learn from other features (e.g. n-grams), emoticons and hashtags are removed from the training data after sentiment labels are assigned. This approach has shown to be successful for SA across various domains and languages as we review in this section.

Go et al. [81]'s work is one of the first to study SA in Twitter using emoticons. They argue that using emoticons as a polarity indicator can show a comparable

performance to sentiment classification of reviews with stars/scores as polarity indicator. Using emoticons, they automatically built a training data-set of 1.6M English tweets. The authors train several machine learning classifiers (SVM, MaxEnt, and NB) to perform binary (positive vs. negative) classification using word n-grams as features. The trained models were then evaluated on a small set of 359 manually annotated tweets with the best accuracy achieved at 83%.

Subsequent work by Bifet and Frank [37] trains classifiers to carry out SA on English tweets using a training data-set that is automatically annotated using emoticons. Unlike Go et al. [81], Bifet and Frank [37] consider experimenting with: balanced vs. unbalanced classes. Training an NB classifier and testing it on the same test-set used by Go et al. [81] yielded accuracy score of 82.45% with the balanced training data. Their experiments on a highly unbalanced data-set (15% negative and 85% positive) yielded an accuracy score of 73.81%. We discuss present techniques to tackle unbalanced classes in chapter 3.

Similarly, Pak and Paroubek [127] used emoticons to collect an English Twitter corpus and build a sentiment classifier. However, they expand the scope of investigations to classify Twitter messages according to their emotional direction to: positive, negative, and neutral. That is, in addition to querying Twitter for positive and negative emoticons to automatically collect the positive and negative instances, they collect a set of neutral instances from Twitter accounts of popular newspapers, e.g. New York Times. Therefore, unlike the binary-classification adopted by Go et al. [81] and Bifet and Frank [37], they experiment with three-way classification (positive vs. negative vs. neutral). Utilising the same test-set used by Go et al. [81], their experiments yielded an F-score around 0.60 with an NB classifier. We follow their idea of collecting neutral Twitter messages from popular news accounts.

Kouloumpis et al. [109] used hashtags to automatically annotate a set of English tweets to experiment on three-way polarity classification (positive vs. negative vs. neutral). They use a set of sentiment-bearing hashtags (e.g. #success, #fail, #job) to automatically build training data and evaluate the trained classifiers on a manually annotated test-set. They report the best results when combining a set of

syntactic, semantic and stylistic features at 74% accuracy and 0.68 F-score.

The approach of DS has also been explored for emotion analysis. Emotion analysis (see section 2.1.1) is concerned about identifying the *type* of emotion being expressed in a piece of text (i.e. happiness, sadness, anger, fear, etc.) rather than its polarity (i.e. positive or negative). In this context, Purver and Battersby [135] empirically investigate the performance of supervised classifiers trained with an automatically labelled training data to perform multi-class emotion analysis. Using emoticons and hashtags, they automatically annotate a set of English tweets into six basic emotion classes which they use to train supervised classifiers (SVMs). The authors deduce that this approach is more suitable for some types of emotions than others. In particular, they find the approach more reliable for detecting happiness and sadness (which corresponds to the binary positive/negative sentiment classification [23]), with the best reported F-scores for evaluating the models on a manually annotated test-set as 0.775 for detecting happiness and 0.545 for sadness with the emoticons data-set and 0.626 for happiness and 0.604 for sadness with the hashtags data-set. It is interesting to see that even with a more fine-grained emotion analysis task, happiness (positive) and sadness (negative) seem to be amongst the most distinguishable emotions.

Similarly, for determining emotion type for less-resourced languages, such as Chinese, Yuan and Purver [180] perform experiments to detect emotions from a Chinese micro-blog service called Sina Weibo, which the authors referred to as the "*Chinese version of Twitter*", and hence characterised by the short length and variety in topics raised and discussed. Emoticons were used to generate emotion labels for six emotion classes. By training SVM classifiers, the authors report that happiness is the most discriminative class with accuracy up to 85.9%. The best reported accuracy for sadness/negative class is at 67.5%. The authors illustrate that happiness and sadness are the most frequent classes among other emotions and speculate that such a pattern is relatively stable across different languages.

As for Arabic, AlMutawa [22] describes a number of experiments the author conducted to carry out emotion analysis on Arabic tweets that were automatically

labelled for six classes of emotion using emoticons. The author utilises syntactic features of word-stem n-grams. By training SVM classifiers on an emoticon-labelled data-set, the author reports accuracy scores up to 57.69% for happiness and 45% for sadness, when evaluating the trained models on a manually annotated test-set. This classifier is beaten by another classifier, which is trained on an automatically labelled data-set using hashtags, attaining accuracy scores up to 63.42% for happiness and 70% for sadness.

We are not aware of previously published work that has addressed the issue of exploiting emoticons and sentiment-bearing hashtags to automatically label Arabic tweets for sentiment polarity. Therefore, in our work described in chapter 5, we conduct a set of experiments to investigate the usefulness of this approach for SA in Arabic tweets.

2.4.3.2 Lexicon-based + Machine Learning

In this approach, researchers explore the feasibility of employing a lexicon-based approach (similar to that described in section 2.4.1) for automatically obtaining sentiment labels, which in turn will be fed into a sentiment classifier as training examples.

Zhang et al. [186] employ a hybrid approach (lexicon-based + ML) to carry out SA on English tweets. First, they apply a lexicon-based approach, which uses a publicly available sentiment lexicon, in order to automatically produce training data. Then, they use the generated examples to train an SVM to perform SA in three-way (positive vs. negative vs. neutral) classification. The best accuracy score is reported at 85.4%.

As for Arabic, El-Makky et al. [64] use a hybrid approach (lexicon-based + ML) to perform SA on Arabic tweets. The authors employ set of word n-grams, semantic, POS, Twitter-specific, stylistic (e.g. presence of elongation) features to train an SVM classifier. Using 10-fold CV, they report F-scores at 0.72 for polar vs. neutral and 0.79 for positive vs. negative. However, the experiments consider only tweets in Egyptian Arabic.

2.4.4 Sentiment Analysis on Arabic Tweets: Issues Identified

In sum, previous work on SA of Arabic tweets suffer from: small data-sets built and investigated on (up to 3k tweets), narrowed feature-sets employed, evaluated in isolation without comparing various approaches and feature-sets against a benchmark test-set to gain a better understanding of how a SA system will perform under different circumstances. More importantly, the classifiers' evaluation in previous work has not considered the dynamic/time-evolving nature of Twitter, i.e. CV or as a split of the original data-set. Some of these questions have already been addressed for English (e.g. [109, 123, 146]). Furthermore, there have been limited investigations into the utility of existing publicly available resources that were originally designed for non-microblogging data in this domain. Finally, there is a need to address the issue of limited annotated corpora available for SA in Arabic, as a less-resourced language, and empirically consider aspects like data quality (manually-annotated) vs. data quantity (automatically-annotated).

2.5 Sentiment Analysis of Arabic Social Media: A Framework

In this thesis, we apply a set of approaches that were found to be successful for SA in the literature and conduct empirical investigations, which include the following steps (see figure 2.2) [74]:

- **Data collection**: this stage queries the Twitter's public stream to collect data.
- **Text Cleaning up and pre-processing**: this step aims to tackle informality and noise typically encountered in social media (e.g. by normalising usernames and lengthening).
- **Sentiment annotation**: with currently no annotated corpora available, we utilise three main methods to address this issue: 1) recruiting human annota-

tors to manually annotate data, 2) exploiting existing elements to automatically label data (distant supervision) and 3) leveraging resources from other languages by using machine translation.

- **Feature extraction**: this component is concerned about representing and abstracting the data using feature-sets that can be useful for SA (e.g. syntactic, semantic, stylistic and genre-specific features).

- **Training a sentiment classifier**: using the annotated data-sets, we can train a machine learning system to classify the sentiment orientation of a given text instance. The trained model is then evaluated to ensure that it is able to perform well on previously unseen data. Here, we experiment with the inclusion/exclusion of feature-sets according to their usefulness. This stage also involves an error analysis and manual examinations of data samples to gain insights and draw observations on the findings.

- **Deploying a system for sentiment analysis**: the best preforming SA classifiers yielded out of the previous investigations and analysis can be deployed and ultimately used to benefit from extracted knowledge - *social media intelligence* - to be used in real-world applications.

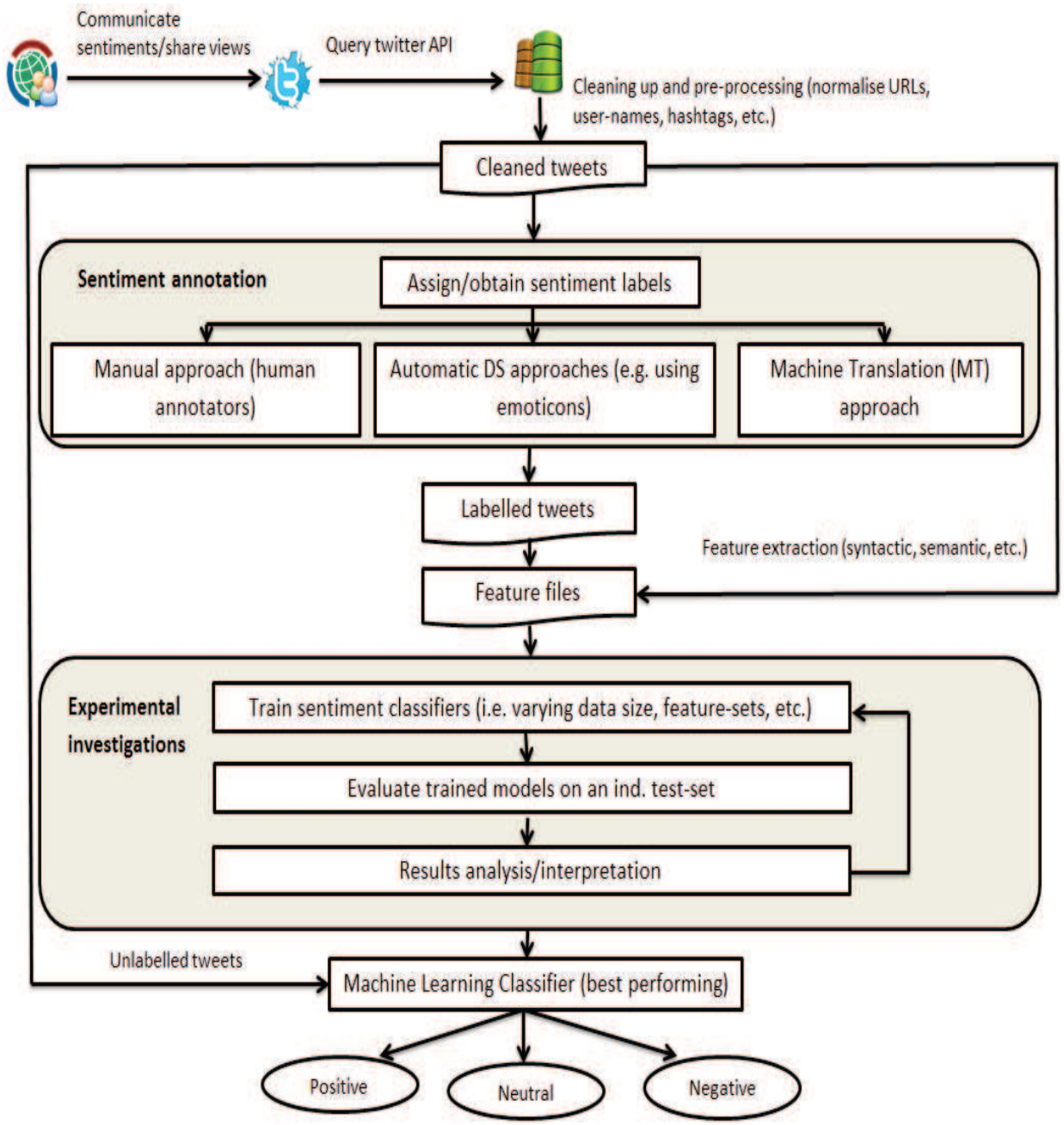


Figure 2.2: A framework for SA of tweets in less-resourced languages.

2.6 Summary

This chapter:

- Explains the main challenges of SA on Arabic micro-blogs.
- Identifies gaps in the current literature.
- Reviews a set of SA approaches that are successfully applied to English.
- Proposes a framework for SA on Arabic micro-blogs, which will be executed in the following chapters.

Chapter 3

Experimental Setup

This chapter provides a description of the creation and annotation of data-sets that we use for the empirical investigations in this thesis. In addition, it outlines a number of pre-processing procedure and feature-sets employed. The chapter also introduces ML schemes and evaluation procedures/metrics employed.

3.1 Data Collection and Annotation

For accessing the Twitter's public stream and collecting Twitter data-sets, we utilise the Twitter Application Programming Interface (API).[1] The API allows Twitter's data to be retrieved by external developers using some search criteria (i.e. keywords, user-names, locations, etc.). Following previous work [81, 186, 77], we search the Twitter API with a pre-prepared list of queries (see table 3.1). Accessing Twitter API is rate limited (180 queries in a 15 minute period), for that, we set a delay/waiting time between requests to be 2-3 minutes, as suggested by Go et al. [81]. To access the Twitter API in a Java-based environment and set queries, we use the Twitter4j Java library,[2] following Fiaidhi et al. [77]. Similar to the work of Purver and Battersby [135] and to avoid bias (i.e. weekends or active users), we collect data on random times of the day and different days of week. In addition, we calculate the distribution of the number of tweets from individual users (using the unique IDs of authors). The recorded rate we observe in our data-sets is between 1.76 to 2.59

[1] `https://dev.twitter.com/`

[2] `http://twitter4j.org/en/index.html`

tweets per user showing no skew towards a group of users. To restrict the retrieved tweets to Arabic only, we set the language parameter of the API to *lang:ar*.

The tweets are retrieved in the form of JSON objects that incorporate a set of meta-data featuring each individual tweet (as shown in figure 3.1). In addition to the tweet text, each JSON object includes features like: a unique identification number of a tweet (tweet-ID), tweet's creation date/time and, optionally, its geographic location. It also outlines some features that could be useful to be exploited for sentiment classification tasks such as: whether a tweet has been re-tweeted or favourited by other users and whether a tweet includes a hashtag, URL or mentioning other users.

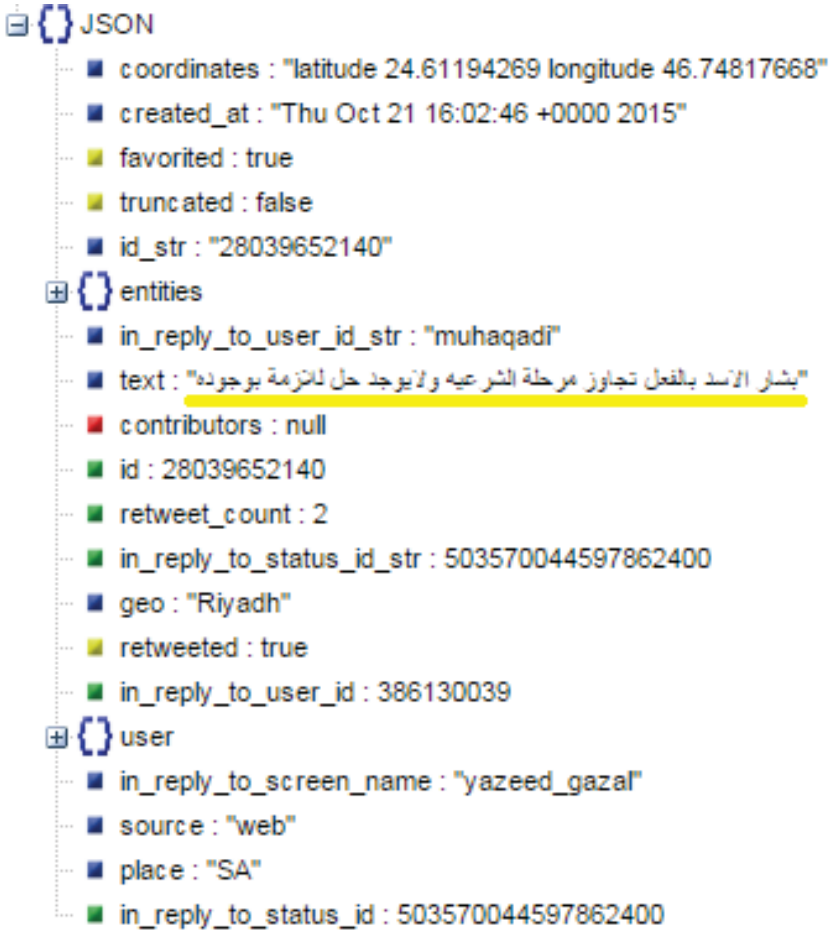


Figure 3.1: An example of JSON tweet.

To account for the issue of redundancy in the Twitter stream (see section 2.2.2 on page 15), re-tweeted and duplicated instances are discarded, following previous work (e.g. [120, 109]). Other noisy content, e.g. ads, are removed from the man-

ually annotated data-sets (section 3.1.1), but they are difficult to filter out from automatically created/annotated data-sets (section 3.1.2), as pointed out by Go et al. [81]. However, we expect that by removing duplicated tweets, such content will be reduced to a considerable extent as we observe that most advertising content tend to appear repeatedly, which is also observed by Dacres et al. [51].

Products/brands	iPhone, Chanel
Social and religious Issues	divorce, education, early/child marriage, Sheia
Public figures	Obama, Mandilla, Khamenei, Erdogan
Sport	Chelsea , Al-Ahli FC
International Committees	United Nations, league of Arab States
Internet and technology	YouTube, Instagram, Google
Controversial topics	Isis

Table 3.1: Examples of query-terms used for collecting the Arabic Twitter Corpus.

For all data-sets described in this section, we did not impose any restrictions on the number of instances of any of the classes, following SemEval [146, 145]. This is likely to obtain a representative sample of the Twitter stream, as highlighted by Bifet and Frank [37]. For instance, one of the SemEval's (2015) Twitter data-sets has a majority positive class with 1,040 instances, while negative class has only 365 instances and neutral class has 987 instances [145].

3.1.1 Gold-Standard Training Data-sets: Manual Annotation

This section describes the collection and annotation of two gold-standard, manually annotated data-sets of Arabic tweets. In order to retrieve tweets which are relevant for SA, we create a set of search queries (as those shown in table 3.1) to increase the chances of obtaining tweets that convey opinions, attitudes or emotions towards the specified entities, following [81, 186, 77, 17, 13, 2, 145]. Note that for training a classifier, these query terms are replaced by placeholders to avoid bias. In particular, we harvested two data-sets at two different time steps:

Gold-Standard data (GS1): This data-set contains a random subset of 2,287

manually annotated multi-dialectal Arabic tweets (out of 15,766 tweets originally retrieved over the period from 25th of January to 5th of March 2013).

Gold-Standard data (GS2): This data-set contains a random sample of 6,894 manually annotated instances (out of 42,247 tweets originally retrieved during the period of 20th of January to 21st of February 2014).

In addition, we use a previously collected corpus:

Mourad and Darwish's data (M&D): This data-set is collected and manually annotated by Mourad and Darwish [120]. The authors have shared this data-set with us and we used it to assess the effectiveness of our feature-sets in comparison to performance scores originally reported by the authors on this data-set.

3.1.1.1 Sentiment Annotation

The two newly collected data-sets (i.e. GS1 and GS2) are manually annotated by two native speakers of Arabic. At the semantic level, tweets can be composed of positive and negative emotions simultaneously [167]. Consequently, this can pose a challenge to the task of assigning sentiment labels, even for a human annotator [73]. This requires guidelines to the annotating scheme and clear definitions of the assigned labels to reduce the chances of overlapping among the defined classes and ensure consistency among annotators (see table 3.2) [172].

Each data instance (tweet) is marked with only a single label that denotes the interpretation that is ultimately conveyed by the complete piece of text, taking into account only the writer's perspectives: *neutral/objective*, *mixed*, *positive* and *negative*, where the latter three are all subsumed under the label *polar*, i.e. subjective, see table 3.2.

The label *mixed* covers the cases where tweets are composed of positive and negative emotions simultaneously [112]. Handling instances with mixed emotions is not trivial. For instance, Read [136] has considered instances with mixed content as noisy data that needs to be removed from the data-set. Pang and Lee [128], on the

other hand, suggest considering a mixture of positive and negative expressions as an overall neutral opinion. In our work, we follow settings used in SemEval's tasks of tweets SA [123, 146] in which mixed instances are assigned a sentiment label based on the strongest emotions expressed and whenever it is difficult to decide, the instance is assigned with a *mixed* label. Abdul-Mageed and Diab [4] report a negligible presence for mixed instances in their Arabic data-sets, while mixed instances represent an average of 6.9% in our GS data-sets. Therefore, we opted to include them under *polar* class as we argue that their ultimate orientation is subjective rather than neutral as suggested by Pang and Lee [128].

Label	**Definition**	**Example**	**English**
positive	Clear positive indicator	كم انت عظيم يَا بشَار الاَسد	*How great you are, Bashar Al-Asad.*
negative	Clear negative indicator	حنَا للّاسف نستخدم ايفون	*Unfortunately, we use the iPhone.*
neutral	• Simple factual statements / news	حَالة وفَاة جديده بَاتش ٧ان ٩ بَالصين	*A new reported death case with H7N9 in China.*
	• Questions with no emotions indicated	بكم سعر الآيفون ٥ حَالياً؟	*What is the price of the iPhone 5 these days?*
mixed	Mixed positive and negative indicators (i.e. difficult to decide on the strongest)	نحن نعشق الديموقرَاطيه و نكره فوضيّ الاخوَان المسلمين التيَ تريد تدمير حريَاتنَا	*We love democracy, but hate the mess that Muslim Brotherhood is making to destroy our freedom*
uncertain	Undeterminable indicators/neither positive or negative/ lack subjective cues	احيَانَا فهمنَا الَّامور بطريقه خطَا يكون هو الصح	*Sometimes, the wrong understanding of things leads to the right thing.*
skip	Redundant or advertising tweets	-	-

Table 3.2: Sentiment labelling criteria for Arabic Twitter Corpus

In cases where annotators are not able to decide on one of the aforementioned labels, they can label tweets with *uncertain*, see examples in table 3.2 from our data-set. Tweets labelled with *uncertain* by at least one of the annotators were excluded from the data-set. The annotators were asked to assign an additional *skip* label to tweets with redundant or advertising content, following Dacres et al. [51].

Agreement study: In order to measure the reliability of the annotations, we conducted an inter-annotator agreement study on the annotated tweets. We use Cohen's Kappa metric [49], as defined in equation 3.1, which measures the degree of agreement among the assigned labels, correcting for agreement by chance. The resulting weighted Kappa reached $\kappa = 0.756$ for GS1 and $\kappa = 0.816$ for GS2, an average of $\kappa = 0.786$, which indicates reliable annotations [45].

$$kappa = \frac{Po - Pe}{1 - Pe} \tag{3.1}$$

$\mathbf{P_o}$: observed agreement.

$\mathbf{P_e}$: probability of chance agreement.

Abdul-Mageed and Diab [4] argue about the difficulty encountered in labelling social media genre for SA, in comparison to newswire for instance. Table 3.3 shows some example annotations from our corpus. For instance, tweets # 1 and # 2, on the one hand, represent cases of an agreement among annotators in labelling tweets with a clear negative polarity, or conveying neutral content. Sarcastic and heterogeneous tweets, on the other hand, have created a challenge even to human annotators, as also noted by Abdulla et al. [9]. Tweet # 3 shows a disagreement among annotators, whether it is a sarcastic view of very complicated and tragic circumstances, or just a negative attitude. In the context of SA, sarcasm is difficult to detect because it uses positive indicators to express negative emotions (e.g. saying *'Oh, what a great day!!'*, while meaning the opposite) [112], see also section 2.2.1 on page 12. To account for the presence of sarcasm, we ask the annotators to - optionally - declare if they think a certain tweet is meant to be sarcastic in the emotions it conveys and use this information to form a new feature (see page 67).

What happens with the examples where both annotators disagree?

A third annotator is employed to decide the selection of the final annotation, if the 3rd annotator disagrees with both annotators, the tweet will be assigned *uncertain* label. Data instances in this category are also excluded from the data-set [30].[3] This procedure is important for the quality of the gold-standard data-set. As

[3]A total of 3,106 tweets are excluded from the Gold-Standard data-sets.

ID	Original tweet	English translation	Anno. 1	Anno. 2
1	لنري قوتكم يَا ارهَابية لنسحقكم و نحن لَا نتشرف بلقيَاكم يَا كلَاب النَاتو	*We will crush your power, you terrorists, and we don't even want to see you, you NATO's dogs.*	Negative	Negative
2	يوجد ايفون بين كل اربعه هوَاتف ذكيه	*There is an iPhone among each of the 4 smart phones.*	Neutral	Neutral
3	علمتنَا الثورَات العربيه ان بشَار الاسد عنده حق	*The political revolution (Arab Spring) has taught us that Bashar Al-Assad is right.*	Uncertain (unclear sentiment indicator)	Negative

Table 3.3: Example annotations from the corpus.

provision of annotated data is a goal of this work, the GS data-set has already been made freely available for the research community via an ELRA repository and by the time of writing this work, the data-set has been accessed more than 162 times and downloaded more than 110 times [138].[4]

As displayed in table 3.8 (page 55), the resultant GS1 data-set has a majority class of neutral instances (representing 50.59%), while positive and negative classes have almost the same number of instances. Again, the GS2 data-set has neutral as the majority class, but with negative instances being almost double the number of positive instances in this data-set. We noted that in manual annotation of Twitter data-set, as the data size increases, negative class seems to be predominantly larger than positive class, which is also reported by Salameh et al. [153] and Nabil et al. [122]. This is probably due to the difficult circumstances in the Arab world in the past few years. The SemEval's manually annotated data-sets of English tweets, in contrast, reflect a clear tendency for more positive tweets than negative ones [146], which is in line with the findings of Dodds et al. [58] who observe that, as social creatures, humans tend to be happier when socialising (i.e. via social media), and hence, their communication (i.e. language used) will generally reflect that positive

[4]Further information about how to access/download the corpus can be found at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

feeling.

3.1.2 Distant Supervision Training Data-sets: Automatic Annotation with Twitter's Conventional Markers

Using the Twitter API and a similar procedure to that described on page 40, we collect a new and much larger training set of Arabic tweets. The only difference here is the set of queries used. That is, instead of querying Twitter for popular figures, brands or controversial social issues, we query Twitter for emoticons or sentiment-bearing hashtags (table 3.4). Emoticons and hashtags are also referred to as *conventional markers* [135]. Again, the data-sets are collected at two different times, with Emo1 data-set is used for investigations on three different SA tasks, and Emo2 and Hash data-sets are used for conducting follow-up investigations on the least performing task, i.e. positive vs. negative (chapter 5).

Conventional Markers Distant-Supervision Arabic data-set (Emo1): This data-set contains a total of 66,471 automatically annotated instances (retrieved during the period of 6th to 15th of November 2013). This data-set is collected and automatically labelled for positive/negative sentiment using emoticons only. We call the resultant data-set Emo1, as displayed in table 3.8.

Conventional Markers Distant-Supervision Arabic data-sets (Emo2 and Hash): The total collected tweets is 2,438,262 (retrieved during the period of August 2014 to April 2015). Theses tweets are collected and automatically labelled for positive/negative sentiment using emoticons and hashtags. After duplicates removal, the resultant data-sets are Emo2 data-set with 1,118,356 tweets and Hash data-set with 130,160 tweets.

Again, we also use a previously collected corpus, but comprises English tweets:

Conventional Markers Distant-Supervision English data-set: This data-set contains a total of 1,600,000 English tweets automatically labelled for posi-

tive/negative sentiments based on the presence of emoticons. We call this data-set Emo-Eng from this point. The data-set is collected by Go et al. [81] and made publicly available.[5] The data-set is balanced (number of positive and negative tweets is equal). As shown in table 3.8, this data-set has no neutral instances. We use this data-set to compare learning rate for sentiment classifiers in English vs. Arabic (see section 5.4.1.2 on page 142).

3.1.2.1 Sentiment Annotation

Following [136, 81, 37, 135], we use a set of emoticons with pre-defined polarity to automatically label training sets of Arabic tweets (i.e. Emo1 and Emo2). As shown in table 3.4, the emoticons we use are frequently used to express positive/negative sentiment [22]. The emoticons are used as noisy labels to serve as an author provided sentiment indicator. Tweets with a positive emoticon will be automatically assigned with a *positive* label; tweets with a negative emoticon will be automatically assigned with a *negative* label; and tweets with mixed emoticons (positive and negative emoticons) are excluded, following Go et al. [81]. Emoticons are merely used to assign the sentiment labels and removed from tweets to avoid any bias and to force classifiers to learn from other features (e.g. word n-grams).

In addition, and following [186, 135], we utilise a set of sentiment-bearing hashtags to query emotional tweets. Similar to emoticons, we use the sentiment-bearing hashtags to collect and automatically annotate a hashtag-based data-set (see table 3.4). Again, hashtags are replaced by placeholders to avoid biasing the data. We call the resultant data-set Hash.

Neutral tweets: In order to collect neutral instances, we query a set of official news accounts, following [127]. Examples of the accounts queried are: BBC-Arabic, Al-Jazeera Arabic, SkyNews Arabia, Reuters Arabic, France24 Arabic, and DW Arabic. Using this method and after duplicates removal, we collected 55,076 neutral instances in total. This auto-obtained neutral set is used with all of the auto-labelled

[5] Available at: https://sites.google.com/site/twittersentimenthelp/home. Accessed on: 09 Sept. 2015

Emoticon	Sentiment label	Hashtag	Sentiment label
:) , :-) , :)), (: , (-: , ((:	positive	(*happy*, سعَاده) (*joy*, بهجـه) (hope, أمل)	positive
:(, :-(, :((, :((,): ,)):)-:	negative	(*sad*, حـزن) (*bane*, مصيبه) (despair,يَأس)	negative

Table 3.4: Emoticons and hashtags used to automatically label the DS-based training data-sets.

DS-based data-sets, i.e. emoticon-based, hashtag-based, and lexicon-based.

The numbers of tweets collected varied in accordance to the popularity of conventional markers (i.e. emoticons and hashtags) that we used to query Twitter. That is, although Emo2 and Hash data-sets were collected over the same period of time, total number tweets retrieved using emoticons is 1,511,621 tweets, while the number of tweets collected using hashtag queries is 926,640 tweets. A similar behaviour was also observed by Purver and Battersby [135] on English tweets. Furthermore, we observe that removing duplicated instances from the emoticon-based and hashtag-based data-sets reveals a very high rate of noisy/repeated tweets in the hashtag-based data-set, resulting in reducing the hashtag-based data-set from 926,640 to 130,160 instances (see table 3.8 on page 55). To illustrate, the discarded content represents 85.9% of the originally collected hashtag-based data-set, as compared to 24.1% of the emoticon-based data-set.

Additionally, for the emoticon data-set, tweets with ambiguous markers, are detected and discarded. To illustrate, a common way to directly quote a text in Arabic is by having it enclosed between parenthesis preceded by a colon, like:

Text :(quoted text)

Therefore, the first part of the quotation can be misinterpreted as a negative emoticon, while in fact it is not the case. The total number of tweets detected with a similar pattern is 28,048 tweets. Table 3.8 on page 55 displays the final numbers used in training sets.

As for class distribution in the auto-labelled data-sets, the emoticon-based (Emo1) data-set has a nearly balanced (positive and negative) distribution while Emo2 (a larger data-set than Emo1) has a final number of 660,393 positive (representing 56.28%) and 457,963 (representing 43.72%) negative tweets. As also observed by Bifet and Frank [37] on an emoticon-based data-set of English tweets, the collected data has more positive instances than negative ones. It is also interesting to note that the manually annotated data-set (GS2) has shown a higher tendency for negative, whereas auto-labelled tweets using emoticons tend to show a higher tendency for positive, which is possibly due to noise. In chapter 5, we discuss issues that we observe with positive emoticons in particular being highly noisy (e.g. mistyped or used sarcastically). The Hash data-set has a final number of 59,990 positive (representing 46.09%) and 70,170 negative tweets (representing 53.91%), a nearly balanced-data-set.

3.1.3 Distant Supervision Training Data-sets: Automatic Annotation with Lexicon-based Methods

In this section, we explore an alternative approach for automatically building training sets exploiting subjectivity lexica. To achieve this, we combine three lexica. We first exploit two existing freely available, manually annotated subjectivity lexica: an Arabic subjectivity lexicon, namely ArabSenti [4] and an English subjectivity lexicon, namely MPQA [174]. We automatically translate MPQA using Google Translate, following a similar technique to [120]. The translated lexicon is manually filtered by removing translations with neutral or no clear sentiment indicator.[6] This results in 2,627 translated instances after correction. We then construct a third dialectal lexicon of 489 words that we extracted from an independent Twitter development set and manually annotated for sentiment. All three lexicons were merged into a combined lexicon of 6,958 annotated sentiment words (duplicates removed). Table 3.5 shows the final numbers in the combined subjectivity lexicon we used in

[6]For instance, *the day of judgement* is assigned with a negative label while its Arabic translation is neutral considering the context-independent polarity.

this work.[7]

Negative	Positive	Neutral	Total
2,693	1,775	2,490	6,958

Table 3.5: The number of entries in the merged subjectivity lexicon.

In order to obtain automatic labels for positive and negative instances, we use the Emo1 and Emo2 data-sets (described in section 3.1.2), remove their emoticon-based sentiment labels and follow two different settings for automatic lexicon-based sentiment labelling: 1) lexicon-presence and 2) lexicon-aggregation. All lexicons and tweets are lemmatised using MADAMIRA (page 21).

The lexicon-presence-based setting: This method automatically labels a tweet as a positive instance if it only includes positive lexicon(s) and the same for the negative class. Data instances having mixed positive and negative words or no sentiment words matching entries from the combined lexicon are excluded from the training set. The resultant data-sets are LexPres1 with 23,455 instances and LexPres2 with 415,858 instances (see LexPres1 and LexPres2 data-sets in table 3.8 on page 55).

The lexicon-aggregation-based setting: This method follows a simplified version of the rule-based aggregation approach of Taboada et al. [165] and Thelwall et al. [167]. For each tweet, matched sentiment words are marked with either (+1) or (-1) to incorporate the semantic orientation of individual constituents. The sentiment orientation of the entire tweet is then computed by summing up the sentiment scores of all sentiment words in a given tweet into a single score that automatically determines the label. To illustrate, the sentiment of a tweet is automatically determined based on the sign of the aggregated score: the tweet is negative if the aggregated score is <zero and the tweet is positive if the aggregated score is >zero. Instances where the score equals zero are excluded from the training set as they represent either tweets with no occurrence of sentiment words from the combined lexicon or

[7]Lexicons can be freely download from: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

mixed-sentiment instances with an even number of sentiment words. Table 3.8 (page 55) presents final statistics of the resulting training sets Lex-Aggreg1 with 28,144 tweets and Lex-Aggreg2 with 487,486 tweets. Under the lexicon-aggregation setting, tweets with mixed sentiment will have positive and negative sentiment-bearing words contribute to the overall sentiment score. Consequently, mixed tweets are included in this data-set with a positive label (i.e. if the overall score is positive) and vice versa for negative. This can give the lexicon-aggregation-based data-sets an advantage (i.e. over lexicon-presence ones) of allowing the inclusion of mixed instances, which we explore the impact of their presence in chapter 5.

In both settings (lexicon-presence and lexicon-aggregation), we account for issues regarding possible bias of 1) sentiment-bearing words exploited and 2) negation scope. For the first issue, the identified sentiment words are replaced by place-holders to avoid bias, following [186]. To account for negation, we reverse the polarity (switch negation), following [165] (see section 2.4.1 on page 27). In this context, Taboada et al. [165] mention that the most likely case is when negation will affect the following word. As for Arabic, Shlonsky [159] points out a similar behaviour for negation clause. However, more variations that can introduce further complexity to the problem are possible in Arabic, e.g. handling negation across dialects [93]. In this work, we consider two common negation variants that are likely to correctly capture the effect of negation to a great extent [159]. These are 1) a negation scope of one token distance and 2) a negation scope of two tokens distance. To illustrate, a negator will have an impact on the polarity of a sentiment-bearing word if it appears maximum within the following two tokens. For instance, examples # 1 (an MSA instance) and # 2 (a DA instance) in table 3.6 show cases wherein negators are instantly followed by sentiment-bearing tokens and in both cases negators will have an impact on the sentiment orientation. Example # 3 represents a case wherein the sentiment-bearing token is two tokens away from negator, but still has an impact on the token's polarity. Example # 4, in contrast, presents an example wherein the distance between the negator and sentiment-bearing tokens increases and hence, will no longer have an impact on the token's polarity. In sum, a negator will reverse

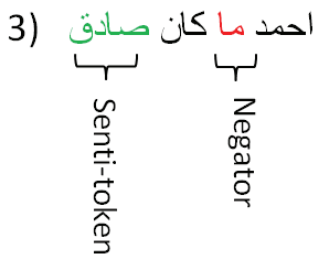
the polarity of a sentiment-bearing word only if it lies within a maximum distance of two tokens following that negator (as in examples 1,2 and 3) and will have no impact otherwise (as in example 4).

1) لم ينجح ريال مدريد في اقناع جمهوره Negator Senti-token *Real Madrid was not successful to convince its audience.*
2) أنا في الشغل اتعامل مع ناس مش محترمة Negator Senti-token *I have to deal with people who are not respectable.*

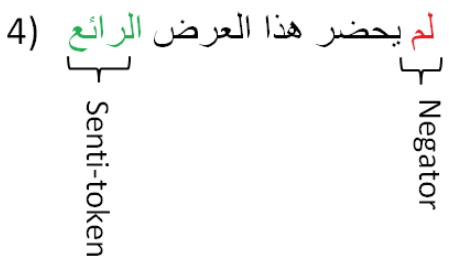
Table 3.6: Examples of tweets with negator instantly followed by a sentiment token.

The class distribution in the lexicon-based data-sets (Lex-Pres1, Lex-Pres2, Lex-Aggreg1 and Lex-Aggreg2) displayed in table 3.8 (page 55) shows positive as the majority class. This indicates that more Arabic tweets in the Twitter stream are having positive words. A recent study reveals that, compared to negative expressions, words that convey positive sentiments are more prevalent and more diversely used in social communications (e.g. Facebook and Twitter) [58].

Finally, the mixed class is empty with the auto-labelled data-sets (as shown in table 3.8) due to one of three reasons: 1) the mixed instances are excluded (i.e. emoticon and lexicon-presence-based data-sets), 2) an external data-set with no mixed instances (i.e. Emo-Eng data-set) or 3) the mixed instances are included under (positive or negative class), depending on the overall aggregated sentiment score (i.e. lexicon-aggregation-based data-sets).



Ahmad is not being honest.



He did not come to this splendid show.

Table 3.7: Examples of tweets with negator followed by a sentiment token at different distances.

Development Data. A sample of 2k tweets were randomly collected and manually annotated for configuration optimisation. The purpose of this data-set is to evaluate the performance of different experimental configurations (section 3.7).

3.1.4 Test Data-set

In order to compare SA systems trained on different training sets, we use an independent test-set to evaluate their performance.

Considering the evolving nature of the Twitter stream (see section 2.2.2 on page 15), we built a test-set that is a collection of random samples retrieved over different periods of time (table 3.9). In addition, the size of the data-set (as shown in the table 3.9) is comparable to that created and used in SemEval on English tweets (sizes for Twitter test-sets are 4,435 tweets in 2013 and 2,473 tweets in 2014). Previous studies on Arabic tweets, in contrast, have considered test-sets that are subsets of the original data-set (e.g. [8]) or used cross-validation (e.g. [120]). Both settings are problematic for Twitter due to its evolving nature and topic-shift issues that are

Data-set	Neutral	Polar *	Positive	Negative	Mixed	Total
Gold standard (GS1)	1,157	1,130	470	467	193	2,287
Gold standard (GS2)	3,697	3,197	876	1,941	380	6,894
Gold standard (GS1+GS2)	4,854	4,327	1,346	2,408	573	9,181
Mourad and Darwish (M&D) [120]	682	1,299	734	377	159	1,981
Emoticon-based (Emo1)	55,076	66,471	32,842	33,629	-	121,547
Emoticon-based (Emo2)	55,076	1,118,356	660,393	457,963	-	1,173,432
Emoticon-English (Emo-Eng) [81]	-	1,600,000	800,000	800,000	-	1,600,000
Hashtag-based (Hash)	55,076	130,160	59,990	70,170	-	185,236
Lexicon-Presence (Lex-Pres1)	55,076	23,455	18,442	5,013	-	78,531
Lexicon-Presence (Lex-Pres2)	55,076	415,858	301,074	114,784	-	470,934
Lexicon-Aggregation (Lex-Aggreg1)	55,076	28,144	18,105	10,039	-	83,220
Lexicon-Aggregation (Lex-Aggreg2)	55,076	487,486	338,765	148,721	-	542,562

Table 3.8: Sentiment label distribution of the training data-sets: gold standard manually annotated and distant supervision data-sets (* Polar = positive + negative + mixed).

likely to influence the predictive ability of a trained model over different points in time (further discussion in chapter 4).

The test-set is manually annotated by two native speakers of Arabic, following the criteria presented in table 3.2 (page 44). The inter-annotator score for the test-set is at $\kappa = 0.69$, which indicates reliable annotations [45]. Our test-set is designed to provide a common ground to build and evaluate SA systems, as it 1) is built with a coverage that spans an extended period of time (see table 3.9); 2) contains less bias to active users (observed distribution of the number of tweets from individual users is 1.16 tweet per user); 3) is annotated with a rich set of morphological, semantic, and stylistic features; and more importantly, 4) is publicly available.[8]

The class distribution in the test-set indicates the negative class as the majority class. This is in line with our previous manual annotations of the gold-standard training data. Following SemEval [146, 145], the instances were randomly selected for manual annotation, which is likely to obtain a representative sample of the Twitter stream [37].

Data-set	Collection Time	Neutral	Polar	Positive	Negative	Total
Test-Sample1	spring 2013	324	377	69	308	701
Test-Sample2	autumn 2013	480	621	285	336	1,101
Test-Sample3	winter 2014	333	518	169	349	851
Test-Sample4	summer 2014	218	667	208	459	885
Total	-	1,355	2,183	731	1,452	**3,538**

Table 3.9: Sentiment label distribution of the test data-set.

3.2 Data Pre-processing

We adapt pre-processing techniques to tackle informality and alleviate the noise typically encountered in social media. We use pre-processing techniques that have been previously employed and shown to be useful for improving performances of SA systems [81, 37, 186, 13, 34, 109, 12, 135, 120, 65, 28, 27, 146]. In particular, the extracted data is cleaned up in a computationally-motivated (i.e. reducing feature space) pre-processing step by:

[8] `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`.

- **Normalising conventional symbols of Twitter:** this involves detecting entities like: #hash-tags, @user-names, re-tweet (RT), and URLs; and replacing them by place-holders.
- **Normalising exchangeable Arabic letters**: mapping letters with various forms (i.e. *alef* and *yaa*) to their representative character.
- **Eliminating non-Arabic characters.**
- **Removing punctuations and normalising digits.**
- **Removing stop words:** this involves eliminating some frequent word tokens that are less likely to have a role in class prediction (e.g. prepositions). For this purpose, we use a publicly available Arabic stop word list that is created by Attia [26].
- **Reducing emphasised words/expressive lengthening**: this involves normalising word-lengthening effects. In particular, a word that has a letter repeated subsequently more than two times will be reduced to two (e.g. *sadddd* is reduced to *sadd*).

Other text pre-processing steps involve:

Text segmentation: This step is performed to separate tokens based on spaces and punctuation marks. For this, we use the publicly available tokeniser called TOKAN integrated into MADAMIRA (page 21) [131].

Text stemming: This is one step further in text pre-processing that aims at alleviating the high dimensionality of the text data by using reduced forms of words (e.g. stems). Abdul-Mageed et al. [8] argue about the importance of employing such a technique, and in particular, when dealing with a morphologically rich and highly derivative language like Arabic, as the problem of high dimensionality becomes more pronounced (see section 2.3.3 on page 24). In this context, Abdul-Mageed [3] highlights the significance of this text pre-processing step and argues that SA on Arabic can be problematic without using the compressed forms of words, as it

will result in the sentiment classifiers being exposed to a large number of previously unseen features (words), although they might be present in training and testing but in different forms. For instance, the words:

وبتَالقهَا *and+with+her+brilliance*, وبتَالقه *and+with+his+brilliance*, بتَالقه *with+his+brilliance* and بتَالقهَا *with+her+brilliance* can be reduced to the stem بتَالق meaning *brilliantly/brightly.*

In sum, stemming has shown to be beneficial for SA on Arabic newswire, reviews and social media posts [7, 19, 13].

3.2.1 Stemming Experiments

This section presents empirical investigations on which stemming type/tool to use. Stemming can be further broken down with respect to the amount of reduction applied to a word into: *root-stemming* and *light-stemming.* The root stemming, on one hand, collapses distinct forms of the words into a representative root, which is typically a sequence of three or four consonants that signifies an abstract meaning of all of its derivations [83]. One of the most common root-stemmers for Arabic is the Khoja Stemmer.[9] Light stemming, on the other hand, can be used to enhance feature reduction while maintaining the meanings of the words. This is performed by reducing the common affixes from the word, instead of reducing the words to their roots. Said et al. [151] argue that using light stemming can enhance the performance of Arabic text classification tasks in general. An empirical question here is: which stemming approach could be more beneficial for SA on Arabic social media text? Previous work has shown preference for the light-stemming over root-stemming [72, 63, 9, 13, 120]. An interesting observation by Farra et al. [72] is that the root-stem can have a negative impact on SA of Arabic as it may reduce a sentiment-bearing word like:

[9] `http://zeus.cs.pacificu.edu/shereen/research.htm`

جميل meaning *beautiful* into a neutral root جمل meaning *a camel*.

To better understand the amount of reduction resulting from each approach, we investigate two publicly available implementations Khoja Stemmer (root-stemmer) and Arabic Light Stemmer (light-stemmer) [150], and run a set of preliminary experiments on the development set (see page 53). Results presented in table 3.10 for binary classification (positive vs. negative) show that the Arabic Light Stemmer can significantly (paired t-test, $P<0.05$) outperform a null-stemmer in which a simple normalisation is performed (i.e. exchangeable letters). The Light Stemmer can also outperform the Khoja root-stemmer but the difference is not statistically significant (paired t-test, $P>0.05$). The number of features generated by each stemmer and used to build the classifier suggest that the Light Stemmer is able to maintain a reasonable balance and trade-off between the sparseness of null-stemmer, and the aggressive reduction of root stemmer that might have caused some sentiment distinctions to be removed (i.e. converted to a neutral stem). Therefore, we use the Light Stemmer setting for all experiments presented in this work.

Stemmer	**No of Feat.**	**Acc.**	**F-score**
Null-Stemmer	30,906	77.56	0.76
Arabic Light Stemmer	19,648	**80.44**	**0.80**
Khoja Arabic Stemmer	10,499	78.10	0.78

Table 3.10: Comparing performances of different stemmers on Arabic tweets.

Stem vs. other word forms: In addition, we investigated the performance of stem against other reduced word forms including: lexeme and tokenised forms. The *lexeme* is obtained by mapping the words to their citation form; while *tokenised* is similar to stem with one exception is that the tokenised morphemes are kept and used as individual features besides the stem. Our experiments show no clear superiority of a single form of word tokens over the others (see table 3.11), with stem being slightly lower than lexeme and marginally better than tokenised (paired t-test, $P>0.05$). Abdul-Mageed et al. [7] find stem to outperform lexeme for SA on Arabic news. In general, stem can be more suitable for the context of language used in social media (i.e. a mixture of DAs and MSA) as stem merely segments clitics with no further processing to the base, whereas lexeme splits off clitics of a

if MADAMIRA encounters a DA word, it will map it to the closest MSA lexeme, which can change its semantic meaning. For instance, we found that the word تخسي, which is a dialectal verb that is used to turn someone/something down in a rude way, is mapped with MADAMIRA into خس lexeme meaning *reduce*. Note that a lexeme should maintain the 'core meaning' of the word it represents [83]. In sum, we decided to use the stem word form for all experiments reported in this thesis as it keeps the base of words, after removing affixes, without further processing and stem is used for SA in previous work [72, 63, 9, 13, 120].[10]

Word form	**Acc.**	**F-score**
Word-stem	80.44	0.80
Word-lemma	80.83	0.81
Word-tokenised	80.23	0.80

Table 3.11: Comparing performances of different word forms on Arabic tweets.

Building feature vector: Finally, the cleaned text instances are passed to this stage of pre-processing in which a feature vector representation of text instances is created [127, 109]. To perform this, we use text pre-processing filters implemented in WEKA [175].[11] Each text instance is represented as a feature/term-weight vector (i.e. in a vector space) where each word token corresponds to a feature in the space (see figure 3.2).

Two major weighting schemes are available for feature vector: *features count* and *features presence*. In this work we use the feature presence that produces a binary value showing the occurrence of a feature regardless of the number times a feature occurs. The choice of feature presence is motivated by the previous work of [129, 70] which was also followed by many studies, e.g. [37, 127, 186, 28] on English, [7, 8] on Arabic and on multi-lingual SA [30].

[10] Overall, we find the effect of applying all of the described pre-processing techniques is resulting an average reduction rate of 64.63% in token frequencies of processed-tweets as compared to the corresponding raw-tweets. Go et al. [81] report a reduction rate of 54.15% by only normalising user-names, URLs and repeated letters.

[11] WEKA is a well-known java-based open-source package that incorporates implementations for a collection of machine learning algorithms for data mining. WEKA is developed by the Machine Learning Group at the University of Waikato, accessing and downloading WEKA is available at: `http://www.cs.waikato.ac.nz/ml/weka/`. In this work we use version 3.7.9 of WEKA.

As shown in figure 3.2, the string attributes will be decomposed via the text filter into strings of single words (unigrams) or multi-words (n-grams). Whenever a new string is encountered, it will be added to the feature space, i.e. the feature vector size will increase by one. Otherwise, if the feature is already existing in the feature space, the corresponding instance containing this feature will have a numerical value of 1, denoting the feature *presence* in this instance.

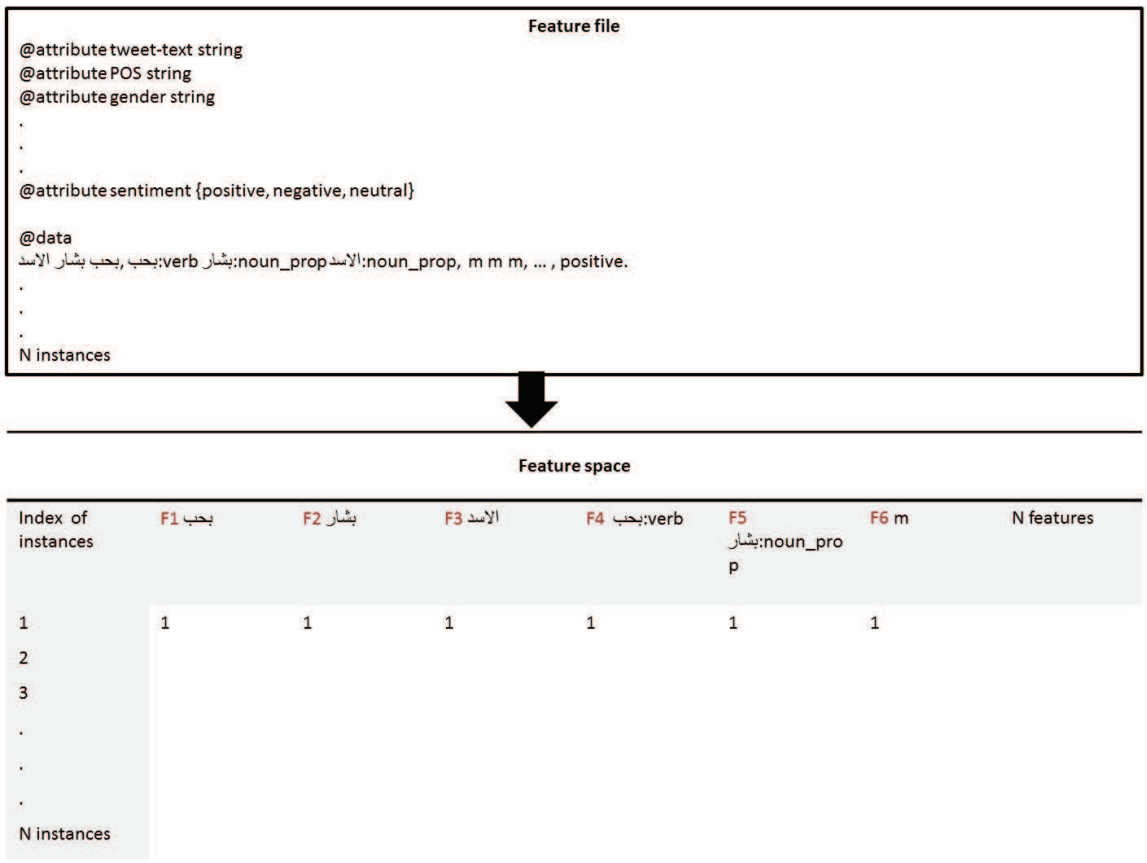


Figure 3.2: An example of the construction of feature space using the feature presence scheme.

3.3 Features Extraction

This section presents a number of feature-sets that we extract and employ to examine their utility for SA on Arabic tweets. The categorisation and design of feature-sets is inspired by the work of Abbasi et al. [1]. Table 3.12 summarises all features we exploit in our investigations.

Word-token-based features: This set involves word-stem n-grams being exploited as features (n-gram sizes are discussed in section 3.7).

Morphological features: The use of this feature-set is motivated by the rich morphology of Arabic (page 24), thus aiming to exploit this aspect by extracting a rich set of morphological features. For that, we employ a state-of-the-art morphological analyser for Arabic, namely MADAMIRA (see page 21) [131]. MADAMIRA on a gold annotated blind test data by Pasha et al. [131] has achieved an accuracy of up to 95.9% for POS tagging and 84.1% for word-level morphological analysis on MSA, as reported by Pasha et al. [131]. As for DAs, MADAMIRA is reported to achieve an accuracy of 63.79% for word-level morphological analysis on Palestinian Arabic (part of Levantine Arabic), as reported by Jarrar et al. [101].

Previous work on Arabic SA has exploited Arabic's morphology in various ways. For instance, Farra et al. [72] manually annotate a small set of movie reviews with a limited set of POS tags and observe an improvement in performance with this feature-set. Additionally, Abdul-Mageed et al. [7] employ a set of six automatically extracted morphological features: person, state, gender, tense, aspect, and number. Although they did not use POS, the authors observe a positive impact with this feature-set on the overall performance on a collection of newswire documents. In [8, 120], the authors experiment on social media data and use a set of automatically extracted POS (using AMIRA, a previous version of MADAMIRA). The authors deduce that the addition of morphological features are beneficial for SA. In our work, we extract a rich set of morphological features that, to the best of our knowledge, has not been previously used for SA on Arabic tweets. Again, this feature-set can be language dependent. In [81, 109], the authors note that adding POS in their SA

Feature-set	Features	Feature type
Syntactic	Word-stem n-grams	String
Morphological	Aspect	String
	Gender	String
	Mood	String
	Number	String
	person	String
	POS:word	String
	State	String
	Voice	String
	Diacritics	String
	Has-morph-analysis	Binary
Semantic	Has-positive-lex.	Binary
	Positive-lex-count	Numerical
	Has-negative-lex.	Binary
	Negative-lex-count	Numerical
	Has-neutral-lex.	Binary
	Neutral-lex-count	Numerical
	Has-negator	Binary
Affective-Cues	Has-consent	Binary
	Has-dazzle	Binary
	Has-laughs	Binary
	Has-regret	Binary
	Has-prayer	Binary
	Has-sigh	Binary
Tweet-topic	Tweet topic	Nominal
Language-style	Tweet-length (char)	Numerical
	Word-length (char)	Numerical
	Word-offset (char)	Numerical
	Has-exclamation-mark	Binary
	Exclamation-mark-count	Numerical
	Has-question-mark	Binary
	Question-mark-count	Numerical
	Has-dots	Binary
	Dots-count	Numerical
	Has-lengthening	Binary
	Has-positive-emoticon	Binary
	Has-negative-emoticon	Binary
	is-Sarcastic	Binary
	MSA-or-DA	Binary
	Degree of dialectness	Numerical
Twitter-Specific	is-Favourite	Binary
	Favourite-count	Numerical
	is-Retweet	Binary
	Retweet-count	Numerical
	Has-hashtag	Binary
	Has-URL	Binary
	Has-user-name	Binary

Table 3.12: A summary of feature-sets used.

experiments on English tweets drop performance.

Morphological feature-set in our work consists of ten word-level features as displayed in table 3.13.[12] It is important to note that the current freely-available release of MADAMIRA is developed for MSA only (see page 21). Tweets, in contrast, contain dialectal and/or misspelled words where the analyser is incapable of generating morphological interpretations. We therefore include an additional feature, namely *has-morph-analysis*. That is, the morphological features for DA words are expected to be noisy. As such, in the following chapters we will be exploring their usefulness, despite noise, for sentiment classification.

Feature	**Example values**	**No of possible values**
Aspect	Imperfective, perfective, N/A	4
Gender	Feminine, masculine, N/A	3
Mood	Indicative, jussive, N/A	5
Number	singular, plural, dual	5
Person	1st, 2nd, 3rd	4
POS	noun, adj, pron, prep	35
State	Indefinite, definite, N/A	5
Voice	Active, passive, N/A	4
Diacritics	Fatha, Damma, Kasra	9

Table 3.13: Morphological features extracted using MADAMIRA.

Semantic features: This feature-set includes a number of binary and numeric features that check the presence and number of occurrence of sentiment-bearing words in each given tweet (table 3.12). To extract this feature-set, we utilise the combined sentiment lexicon described in table 3.5 on page 51. Our merged sentiment lexicon exhibits a reasonable degree of coverage/variation as the translated and filtered MPQA and ArabSenti are expected to represent more formal language (both are in MSA), while our Twitter-based lexicon is expected to represent informal and dialectal language, contributing words like:

طز *go to hell* and بلطجي *bully*.

[12]Further details about the definition of each feature can be found in MADAMIRA's user manual [15].

Additionally, this set has a feature to indicate the occurrence of negators, as n-grams can be inadequate for capturing negation, in particular when occurring at longer distances from polarity words [165].

The semantic features have shown to be beneficial for SA on English tweets [12] and Arabic newswire [7] and movie-reviews [72]. However, Abdul-Mageed et al. [8] observe that a set of semantic features, extracted using ArabSenti [7], to have no impact on SA of a data-set of 3k Arabic tweets. In this work, we exploit the same subjectivity lexicon and explore the impact of expanding it to adapt to the domain of social media (i.e. by including dialectal sentiment-bearing words).

Affective-Cues/Social-Signals: This feature-set comprises six binary features (as displayed in table 3.15), indicating whether a tweet has any of these social signals: consent, dazzle, laughs, regret, prayer, and sigh. To obtain these features, we use six manually created dictionaries.[13] To avoid bias, the extracted dictionaries are based on an independent data-set that does not overlap with any of data-sets described in section 3.1.

The use of this feature-set is motivated by the idea of finding a set of simple features that can correlate to users' culture and, at the same time, can be used as a means for conveying sentiments. The work of Ptaszynski et al. [134] has inspired this idea, as they employ a manually collected lexicon of emotive expression, i.e. culturally-specific Japanese emotional expressions, and note that these features are useful for SA on Japanese blogs. As for Arabic, we find interesting observations reported in previous work on the use of culturally-specific expressions to convey sentiments. For instance, Mourad and Darwish [120] observe a tendency of users to express their feelings through extensive use of Quranic verses and Prophetic sayings, i.e. religious related. Similarly in [19, 64], the authors report that Arabic users tend to use popular compound phrases and idioms to express their sentiments. Table 3.14 shows an example of a religious quote that carries a prayer and usually used to convey a negative attitude. We are not aware of previous attempts to utilise similar

[13]The lists are freely available at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

features for Arabic SA.

حسبنَا اللّه ونعم الوكيل
seeking Allah for a revenge

Table 3.14: An example of an affective cue (prayer) typically used to convey sentiment.

Feature	**Example words**
Has-consent	(اكيد, *sure*) (بَالفعل, *definitely*)
Has-dazzle	(وَاو, *wow*) (مدهش, *impressive*)
Has-laughs	(هّه, *hh*) (لول, *lol*)
Has-regret	(اسف, *sorry*) (حسَافه, *regret*)
Has-prayer	(اللّهم, *O'God*) (حسبنَا اللّه, *O'Allah*)
Has-sigh	(اه, *sigh*) (يَاَاه, *sigh*)

Table 3.15: Affective Cues features along with examples of words used to determine the value of each feature.

Tweet topic feature: This feature aims to incorporate topic information into a feature and assess its role in SA. Abdul-Mageed et al. [7] employ a manually assigned topic feature indicating the domain each sentence is representing, e.g. politics or sports, and report this feature to be useful for SA on Arabic news. Following Abdul-Mageed et al. [7], we ask the annotators to select one out of a set of pre-determined broad topics, which are: sport, economy, commercial, politics, social/religious, internet, and other. The aim is to study the correlation between the topic being discussed and sentiment conveyed, e.g. whether users tend to have negative attitudes when discussing political issues. This feature is only associated with the GS data-sets since it is manually assigned. Future investigations can involve automatic extraction of this feature-set, e.g. topic modelling [111].

Language style features: This set involves a number of features that characterise the language typically used in social media, including:

A) **Stylistic features:** This set of features is also referred to as language independent. It captures information about the informal language used in social media

and may convey sentiment. That is, stylistic features aim to unveil "latent patterns" that can improve classification performance of sentiments [1]. This set comprises features checking for stylistic variation, i.e. presence of: emoticons, expressive lengthening (e.g. *sadddd*)[14] and ungrammatical use of punctuations. In addition, stylistic feature-set incorporates quantitative features like: tweet length (char), word-length (char) and word-offset, which is calculated as char distance from the beginning of the tweet to the first char of the corresponding word. We expect this set of stylistic features to be beneficial, especially in social web, which is rich in such stylistic variation [1]. For instance, Kouloumpis et al. [109] and Thelwall et al. [167] found stylistic features to be amongst the most informative features for SA on English tweets. In addition, Abbasi et al. [1] reported this feature-set to be helpful for SA on English and Arabic forums. Mourad and Darwish [120] used a set of stylistic features for SA on Arabic tweets, but did not report on the impact of this feature-set.

B) **is-Sarcastic Feature:** This is a binary feature assigned by the human annotators, who also assigned the sentiment labels, to declare if they think that the intended sentiment of a certain tweet is being conveyed sarcastically. The selection of a value for this feature is optional, so that annotators will set *is-sarcastic:true* if they believe a tweet is involving sarcasm, otherwise, the feature's default value is *is-sarcastic:false*. Because this feature is manually assigned, it is only used with the GS data-sets. Automatic sarcasm detection is a research area of its own that has received a considerable attention [32, 112], and addressing this issue is beyond the scope of this work. However, due to its potential impact on the sentiment orientation of a text instance, we assess the impact of employing is-Sarcastic feature with the SL experiments (chapter 4). As for the DS experiments, wherein the sentiment labels are assigned automatically, we consider the impact of potentially sarcastic tweets on the overall performance by manually examining data samples for sarcasm as a part of error analysis (see page 138). In SemEval [146], for instance, they have accounted for sarcasm by

[14]This pattern is normalised in the data pre-processing stage (see page 56), but we utilise a binary feature that accounts for the presence of this pattern *has-lengthening: true,false*.

building a small set of 86 tweets containing #sarcasm hashtag on which they observe a poor performance for SA systems. Overall, the percentage of tweets with *is-sarcastic:true* in the GS data-sets is 3.27%.

C) **MSA-or-DA feature:** This is a binary feature to investigate the usefulness of employing an explicit feature that identifies the language variety of a tweet instance (MSA or DA). To automatically extract this feature, we use AIDA (see page 22) [66]. In addition to identifying the language variety of a tweet as MSA or DA, AIDA can provide a numerical value between [0,1] reflecting the degree of dialectness for the corresponding tweet, which we also exploit as a feature.

MSA-or-DA feature can be particularly useful for investigations on Arabic tweets to assess the impact of DA presence on the overall performance of SA. The use of this feature is also motivated by the fact that MSA is often referred to as "*the language of the mind*" while the DAs as "*the language of the heart*" [56].[15] This feature has been used by Abdul-Mageed et al. [8] who report no significant gain for SA. In this work, we do not use this feature on its own, instead we combine it with other language-style features, including degree-of-dialectness, and examine its impact on a larger data-set than that used by Abdul-Mageed et al. [8].

Twitter-specific features: This set utilises seven features characterising the way Twitter is being used (table 3.12 on page 63). Twitter can be used in various ways: for information sharing (via inclusion of URLs and hashtags) and/or for social networking (via inclusion of user-mentions and re-tweets), such uses vary across languages [92]. For instance, Hong et al. [92] investigated behaviour differences among users of different languages and observed that communities like Korean and Indonesian tend to exhibit more for social networking, whereas English and German users tend to use Twitter more for information sharing. We are not aware of a similar study on Arabic. Thus, we explored one of our own data-sets comprising

[15]For instance, we find that Dialectal tweets represent 34.12% of the negative tweets, 37.39% of the positive tweets, and only 13.52% of neutral tweets in the GS data-sets, suggesting subjective instances to be more dialectal as compared to neutral ones. In addition, Cotterell and Callison-Burch [50] reported 40% of their Arabic Twitter data-set comprising >40k tweets were manually annotated as highly dialectal.

1.2M Arabic tweets (i.e. our Emo2 data-set, see table 3.8 on page 55) and observed a higher tendency for social networking (e.g. up to 36.80% of tweets included user-mentions), while only an average of 16.64 % of tweets included hashtags/URLs, i.e. less use of tweets for information sharing. In this work, we employ these means as features including: the presence of hashtags, user-mentions and URLs. In this context, Thelwall et al. [167] noted an association between positive sentiment and use of URLs that the authors assumed it to be a result of people tending to provide URLs in a context of recommendations with positive statements. Additionally, we use other Twitter-specific meta-data that can be automatically retrieved along with each tweet and that can imply sentiments. For instance, whether a tweet has been favoured or re-tweeted may imply support to a view expressed in the tweet. This feature-set has been used in previous work and found to be useful for SA on English tweets [12, 146]. On the contrary, Mourad and Darwish [120] use this set of features for SA on a set of <2k of Arabic tweets and report them to not be discriminating enough. In this work, we further investigate the utility of this feature-set by assessing their usefulness on a larger data-set.

Conclusion: Feature-sets. The experiments are designed to assess the individual contributions of feature-sets described in section 3.3. Table 3.12 (page 63) summarises all features we exploit in our investigations. With word-based n-grams (e.g. stems) are found to be amongst the most informative features for SA on Arabic tweets [120], we use stem n-grams as a base and then add the feature-sets individually to explore their individual impact on the overall performance, following Agarwal et al. [12]. Finally, we experiment with a combination of all feature-sets.

3.4 Levels of Sentiment Classification

We experiment with two alternative problem formulations for sentiment classification: *two-level binary classification* and *single-level/flat three-way classification*. Related work has treated subjectivity and sentiment analysis as two-stage binary classification process, where the first level distinguishes polar/subjective vs. neu-

tral/objective statements (figure 3.16), and the second level distinguishes polar/-subjective instances into: positive vs. negative (e.g. [120, 8]). Alternatively, the classification can be carried out at as single-level three-way classification positive vs. negative vs. neutral (e.g. [72, 12, 146]). We experiment with both options by collapsing the positive and negative labels into the polar label.

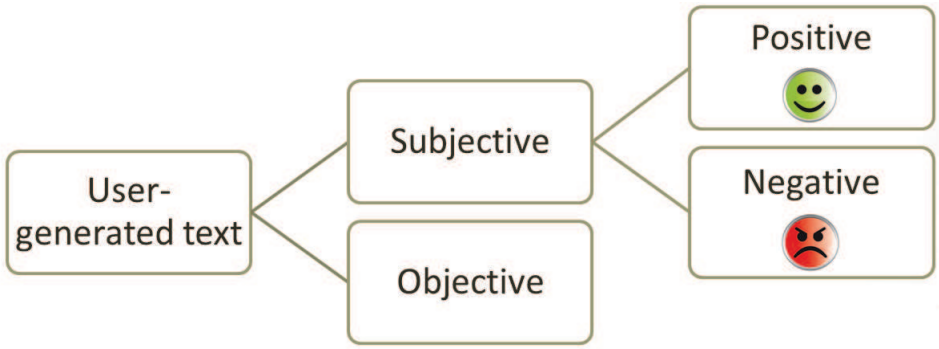


Table 3.16: Levels of Sentiment Classification.

3.5 Machine Learning Schemes

In this work, we use Support Vector Machines (SVMs) [102] as a machine learning scheme that is found to be particularly successful for text classification problems, including SA [129, 30, 146, 145, 3]. This because of their ability to handle a large number of features in a high dimensional feature space (i.e. text classification problems) [102, 94, 103]. In addition, SVMs are an appropriate tool for SA on microblogs tasks due to their ability to be resistant to noise/variance (i.e. L2-regularised solver) [115]. A trained SVM will attempt to classify a new instance to one of the pre-defined classes on which the model is originally trained by finding a hyperplane/decision-surface that separates the instances of classes (figure 3.3) [175]. Two more hyperplanes parallel to the separating hyperplane are created, also called *support hyperplanes*. The support hyperplanes cut through the closest training instances, which also called *support vectors*, on either side [103]. An important characteristic of the supporting hyperplane is that the margin between the hyperplane and the nearest data points on each side is the maximal [94]. To map training data/vectors into a higher (maybe infinite) dimensional space, a kernel function is

used. In this work we use the linear kernel, as suggested by Hsu el al. [94] to be appropriate with text classification problems (i.e. number of features is large).

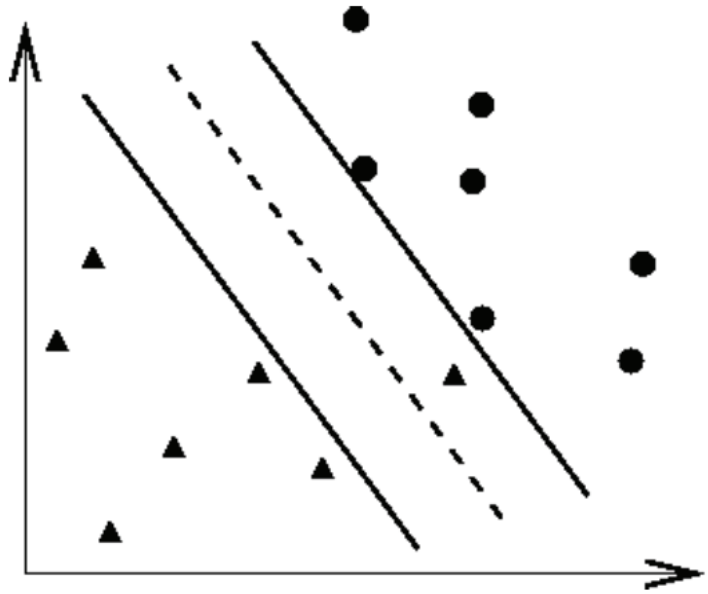


Figure 3.3: An SVM classifier. The triangles and circles represent data instances of two different classes [94].

There are several implementations for SVM that have been successfully used for various text classification tasks. Three prominent implementations are Sequential Minimal Optimisation (SMO) [132], LIBSVM [46] and LIBLINEAR [70]. Our preliminary experiments (see table 3.17) show that the three schemes are able to attain comparable performances, i.e. with no statistical significant differences (paired t-test, $P>0.05$). However, LIBLINEAR and LIBSVM are able to significantly outperform SMO with respect to training time (paired t-test, $P<0.05$). To choose between LIBSVM and LIBLINEAR, we follow guidelines by Hsu et al. [94] who show that LIBLINEAR is more efficient in tackling document/text classification problems - wherein both the number of instances and features are large - than LIBSVM in terms of the time required to obtain a model with a comparable accuracy and memory consumption. Therefore, we use LIBLINEAR for all experiments reported in this work.[16]

[16]LIBLINEAR's implementation is available at: `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`

SVM implementation	**Acc.**	**F-score**	**Training time** (sec.)
LIBLINEAR	79.97	0.796	0.162
SMO	79.49	0.791	8.538
LIBSVM	79.48	0.791	1.354

Table 3.17: Comparing performances of different implementations of SVM.

3.5.1 Baselines

This section outlines a number of baselines that we compare our results against.

Majority Baseline (B-Mjr): Following previous work [7, 8, 146], we compare our results against a majority baseline (i.e. the ZeroR classifier in WEKA) that always predicts the most frequent class in the data-set. ZeroR is a useful classifier to provide a lower bound on the performance of the data-set [175].

MSA Baseline (B-MSA): The aim of this baseline is to assess the performance of a sentiment classifier that is trained only on a set of tweets identified as being written in MSA and evaluate it against a test-set with MSA+DAs instances [185].

Stem n-grams Baseline (B-stem): The purpose of this baseline is to explore whether the addition of individual blocks of feature-sets can result in a significant difference over a sentiment classifier that is trained on stem n-grams features only. A previous study by Agarwal et al. [12] used word n-grams model as a baseline for SA classifiers on tweets.

3.6 Performance Evaluation

This section outlines measures and techniques we adopt to evaluate the performance of our sentiment classifiers.

3.6.1 Evaluation Metrics

In classification problems, the overall performance is measured by identifying the success rate, which is the proportion of the correctly classified instances over the

entire set of instances. We report the results with two metrics: weighted F-score [17] and accuracy.

Accuracy is one of the most widely reported metric in literature and is calculated as:

$$Accuracy = \frac{number\ of\ correctly\ classified\ instances}{total\ number\ of\ instances} \tag{3.2}$$

The *F-score* is defined as the harmonic average of precision and recall[18] and is calculated as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.3}$$

Where *precision* is calculated as:

$$Precision = \frac{A}{A + C} \tag{3.4}$$

A: number of correct/relevant instances classified/retrieved.

C: number of incorrect/irrelevant instances classified/retrieved.

and *recall* is calculated as:

$$Recall = \frac{A}{A + B} \tag{3.5}$$

A: number of correct/relevant instances classified/retrieved.

B: number of correct/relevant instances not classified/retrieved.

3.6.2 Evaluation Methods

Assessing the success rate of a classifier on previously unseen instances - that has played no role in building the classifier - should provide a reliable indicator of the

[17] Following SemEval [146], we use the weighted F-score, which is the average of all f-scores attained for each class (i.e. F-positive, F-negative and F-neutral). That is, each F-score is weighted according to the number of instances with that particular class.

[18] A control parameter β can be used to decide how much emphasis to put on precision vs. recall. F1, or by convention F, is where β's value is 1 denoting an equal/balanced emphasis on both metrics.

classifiers' future performance [175]. We use two options for evaluating the trained models: cross-validation and independent test-set.

3.6.2.1 Cross-Validation (CV)

Cross-validation uses a fixed number of data proportions, namely folds, in order to split the data into test and training sets. The data-set is randomly reordered before being split into n folds of equal sizes. In each fold, every class is represented by approximately the same fraction as in the full data-set, which also called *stratified CV*. Previous work has used 10 as the number of folds to get the best estimate of error [175]. Each fold is then held-out to be used in turn for testing. This makes the learning process run 10 times on different combinations of the training set. At the end, the resultant 10 error rates are averaged to yield the overall score. As an enhancement for reliability of the results, and as suggested by Witten et al. [175], we ran 10 experiments of different 10-fold CV for each data-set; which results in 100 invokes of each learning algorithm on each data-set with scores averaged over 10 repetitions.

3.6.2.2 Independent Test-set

Supervised machine-learning techniques for SA on social media can be sensitive to the degree to which the training and testing data match with respect to topic and time-period [136, 137]. That is, the performance of classifiers would normally drop as mismatch between training and testing data-sets increases, i.e. data-sets are about different topics (topic dependency) or from different time-periods (temporal dependency) [137]. Additionally, the size of the test-set affects the success/error rate estimation [175]. For instance, the performance of one of the top performing SA system on English tweets dropped from 83.0% accuracy on a small data-set of 359 manually annotated tweets in [81] to 66.46% when tested on a data-set that is much larger, more diverse and collected at a later point in time than that of the original training data-set [2]. Previous work on SA of Arabic tweets either used only cross-validation (each fold $\approx$ 200 tweets), as in [120], or used a hold-out test-set that is a subset of the original data-set, as in [8]. We therefore collected and

manually annotated an independent and diverse test-set (see table 3.9 on page 56) to evaluate the models' ability to transfer/generalise in a dynamic medium like Twitter, following SemEval [123, 146, 145].

3.6.3 Statistical Tests

Carrying out statistical tests is needed to provide evidence that variation among different classifiers is not caused by chance [175]. In our case, we need to ensure that models trained on different data-sets (i.e. sentiment labels obtained using different approaches) or models trained on the same data-set, but with different feature representations, can perform *significantly* different. In this work, we use two statistical tests that have been widely used in text classification problems, namely T-test and Chi-squared (χ^2) [175]. Both tests are conducted on accuracy and at a confidence interval of 95% ($p < 0.05$), as in [1, 78].

T-test is a statistical method that is used to measure the difference between the means of two samples. We used this test with the CV wherein each fold of CV yields a different and independent error estimate. The computed means of the obtained error estimates is used to determine if the mean of a sample of error estimates is significantly greater than, or significantly less than the mean of another. The use of t-test with CV settings is because the performance here is represented by a continuous/scale variable of accuracy (a natural number $\in$ [0-100]) for each fold of data [175]. As such, we use t-test with CV setting, following Abbasi et al. [1].

Chi-squared (χ^2) is a popular test that is used with categorical data. We used this test in our experiments on the independent test-set. We use χ^2 on the independent test-set to test if the observed proportions of a categorical variable (e.g. sentiment labels as those in figure 3.4) differ significantly from a known distribution/proportions, i.e. gold-standard labels [80]. χ^2 significance test has also been reported in previous text classification studies, as in [165]. In addition, we report

on a measure for effect size, namely phi ϕ (equation 3.6) [78].

$$\phi = \sqrt[2]{\frac{\chi^2}{N}} \tag{3.6}$$

where N = the number of observations. Finally, we also report on the *classification error*, which accounts for classification errors on the individual predictions (equation 3.7) [42].

$$ClassificationError = \frac{1}{N}\sum_{i=1}^{n}(1 - isCorrect) \tag{3.7}$$

where *isCorrect* is a categorical variable reflecting if a given test instance *is-correctly-predicted: true=1, false=0* (figure 3.4).

```
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 0.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: negative,1, True
ID: 1.0, actual: negative, predicted: negative,1, True
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: negative,1, True
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 0.0, actual: negative, predicted: negative,1, True
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 0.0, actual: negative, predicted: negative,1, True
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 1.0, actual: negative, predicted: positive,0, False
ID: 0.0, actual: negative, predicted: positive,0, False
```

Figure 3.4: A snapshot of an output file showing for each test instance: the actual (gold-standard) label, the label predicted by the trained model and a binary value showing whether the actual and predicted labels are matched.

3.7 Experimental Setting Optimisation

This section presents a number of decisions made based on a set of preliminary experiments regarding aspects such as:

Size of feature vector: The aim is to decide on *how many features/words to keep*. Reducing number of features is mainly computationally-motivated, especially in text classification problems wherein the number of features is usually large [94]. However, this exclusion of less-frequent features should not affect the performance of classifiers. In [7, 8] the authors set different thresholds for minimum frequency (i.e. <5 and <3) but report that it hurts the performance in some cases. Our findings on the development set are in agreement with this, in the sense that setting various threshold values for the minimum-frequency parameter has shown no gain over the default value (minimum frequency is set to 1). Therefore, we have decided to keep all words for the manually-labelled GS data-sets. The reason for that is that the GS data-sets are relatively small as compared to the other data-sets (e.g. DS data-sets); in addition they are manually annotated, hence the expected noise is much less. For the remaining data-sets, which are larger and auto-labelled (i.e. DS data-sets, see table 3.8), we found that the size of feature vector can be as large as 12M. Subsequently, we assessed the performance of classifiers on different sizes of features on a logarithmic scale. We find that the best performance is yielded at a feature vector of 150k features.

Size of n-grams: Prior research has shown a superior performance for unigrams (1g) and combination of unigrams (1g) + bigrams (2g) over higher order n-grams for the task of SA both on English [81, 127, 109, 28] and Arabic [7, 149, 13, 120, 19]. They are also found to outperform higher orders of n-grams on text classification tasks other than SA, e.g. dialect identification [50]. This aspect might be language dependent as Yuan and Purver [180] observe that 2g and 3g can outperform 1g on Chinese micro-blogs text.

Although 1g on their own are found to be informative [129, 81, 7, 13], Pang and Lee [128] argue that 1g can miss contextual information that is valuable for SA (e.g. negation). Therefore, the combination of 1g + 2g can capture contextual information and attain a trade-off between the coverage of 1g and sparsity of higher order n-grams (i.e. 3g) [127, 120], which is found to hurt performance in Arabic SA [7, 13]. Results of our experiments (table 3.18) are in line with findings of prior

research, as we find that, unlike 1g+2g+3g which can be significantly worse than 1g, the combination of 1g+2g is found to be either significantly better than 1g or as good as 1g, but never significantly worse than 1g. Accordingly, we use 1g+2g in all the experiments reported in this work.

n-grams	**Acc.**	**F-score**
1g	79.97	0.80
1g+2g	**81.21**	**0.80**
1g+2g+3g	79.55	0.78

Table 3.18: Comparing performances of different sizes of n-grams on Arabic tweets.

Tuning C parameter of SVMs: C is the penalty parameter of error and the most important parameter to obtain an optimal value for in SVMs with linear classification [94, 70]. C determines how much to avoid misclassification of data instances, i.e. producing a hyperplane that correctly separates as many instances as possible (section 3.5). The default value is C=1 and adjusting this value (increasing or decreasing) is likely to improve the performance of SVMs on new instances. Larger values of C will result in a larger margin separating instances, while smaller C values will result in a smaller margin. Deciding on C value and producing the optimal hyperplane can vary depending on data. In our preliminary experiments on the development set (table 3.19), we tried a wide range of values using CV, as recommended by Hsu et al. [94] and Fan et al. [70]. The C values in table 3.19 are automatically produced by setting a lower-bound, an upper-bound and the number of optimisation steps [107].[19] The choice of optimisation ranges we used is based on guidelines by Fan et al. [70]. Overall, values C>1 have not yielded any improvement; while values C<1 have resulted in a non-significant (paired t-test, P>0.05) gain at C=0.141. Trying even smaller values like C=0.01 and C=0.001 has resulted in a significant drop (paired t-test, P<0.05) in accuracy. Therefore, we opted to experiment with the default value of C=1 in all experiments reported in the following chapters.

[19]We use optimisation tools implemented in WEKA to automatically select C value, see `https://weka.wikispaces.com/Optimizing+parameters`

[lower-bound, upper-bound] no. of steps	selected C	Acc.	F-score
[1, 10] 10	1	80.16	0.79
[0.5, 1] 10	0.555	80.72	0.79
[0.1, 1] 15	0.165	80.79	0.79
[0.1, 0.5] 30	0.141	80.94	0.79

Table 3.19: Comparing performances of different C parameter values on Arabic tweets.

Handling highly unbalanced classes: Although SVMs can handle class variation to a great extent, there is a chance for producing a suboptimal model which is biased towards the majority class and perform less effectively on the minority class [35]. Two major solutions have been widely used in literature to address this issue in classification problems [48, 14], the problem is tackled either:

a) Internally (using class weights);

b) Externally (using over/under sampling techniques).

In the first solution, developers can assign weights for classes which lead to higher misclassification penalties to training instances of the minority class, i.e. setting the classes with weights to the inverse of the imbalanced ratio [14]. For instance, negative class is two times larger in size than positive class in our GS2 data-set (table 3.8), we can then multiply the weight of positive class weight by two. However, this solution has resulted in no significant gain for the overall performance in our experiments.

The second solution has implemented a number of techniques to address this problem. One popular technique is Synthetic Minority Oversampling TEchnique (SMOTE) [48]. This method involves over-sampling the minority class by creating synthetic minority class examples. We explored the impact of applying SMOTE by experimenting on GS2 data-set as it shows unbalanced class distribution. Applying SMOTE has resulted in a significant (paired t-test, $P<0.05$) improvement of up to 7% in accuracy (table 3.20). However, using SMOTE with larger (e.g. DS) data-sets results in significant increase in training time and memory consumption required to obtain synthetic examples [48]. As such, we opted not to use SMOTE with the DS data-sets.

SMOTE	Acc.	F-score
without SMOTE	80.68	0.80
with SMOTE	**87.72**	**0.88**

Table 3.20: Comparing performances of an SA classifier with vs. without SMOTE on Arabic tweets.

Automatic feature selection/reduction: A common procedure in machine learning is feature selection that precedes learning of a classifier. The aim of this process is two-fold: dimension reduction, i.e. speed improvement, and performance improvement by discarding attributes that are irrelevant and hence, can confuse the machine learning classifiers [175]. We experimented with two of the most well-known attribute-selection methods, namely Chi-squared (χ^2) and Information-gain (IG) [175]. χ^2 evaluates features by computing the chi-square statistic with respect to the class. IG evaluates features by measuring their information gain with respect to the class [175]. In all experiments, the use of automatic feature selectors has not yielded a significant gain; on contrary, the performance of the classifiers was hurt in few cases. This was also observed by Thelwall et al. [167]. As a result, we chose not to use any of these methods in our experiments, but we used Chi-square to obtain ranked lists of the most informative features for error analysis purposes and to gain insight about what subset of attributes are beneficial and discriminative.

3.8 Summary

This chapter established the experimental framework that forms the basis for the empirical investigations in the following chapters 4-6. This involves aspects like the collection, pre-processing and annotation of training and testing data-sets. We created several training sets using three main approaches: 1) manually using human annotators, 2) automatically using Twitter's conventional markers and 3) semi-automatically using lexical-based methods. In addition, the chapter presented the features we extracted and the machine learning schemes and baselines we employed. The chapter outlined the evaluation approaches and metrics we use in this work.

Chapter 4

Supervised Learning Approach

This chapter investigates a supervised machine learning approach (SL), exploiting gold-standard data-sets for accurately classifying sentiments of Arabic tweets. SL approaches are amongst the most successful and popular methods used for sentiment analysis on English tweets [146]. Parts of this chapter are published in [141].

4.1 Related Work

Previous work on SA has used manually annotated gold-standard data-sets to analyse which feature-sets and models perform best for this classification task. The most prominent work by far in this area is the popular SemEval tasks for SA on English tweets in its three editions in 2013, 2014 and 2015 [123, 146, 145]. For this task, a benchmark data-set of nearly 10k tweets is created and manually annotated for positive, negative and neutral. The test-sets used were collected at different points in time than that of the training data, allowing for different topics to be covered in training and testing data [145]. We followed this approach when creating our test data (see section 3.1.4 on page 54). SemEval includes a number of sub-tasks, e.g. determining overall polarity, contextual polarity of a phrase, among others. The work presented in this thesis is closely related to sub-task B. In particular, sub-task B aims to classify a given tweet instance into positive, negative or neutral (from its author's perspective). The top performing systems on this sub-task attained F-scores ranging 0.248-0.648 on English tweets [145], with the majority of systems

using supervised learning methods and employing popular machine learning classifiers typically used in text-classification problems, (e.g. support vector machines (SVM) or Naïve Bayes (NB)). In this work, we follow a similar experimental setup and utilise most of the feature-sets explored in SemEval, including: word-based n-grams, Twitter-specific, semantic and stylistics features, exploring their effectiveness for SA on Arabic tweets.

Supervised learning methods have also been successfully used for SA on Arabic newswire, e.g. [7], and reviews, e.g. [72], with scores of up to 95.54% and 84%, respectively. So far, only a few studies have investigated Arabic social media (as summarised in table 4.1). For instance, Abdul-Mageed et al. [8, 6] present a system for SA on Arabic social media content including a Twitter data-set of 3k tweets. Training an SVM, the best results on the Twitter data-set are reported at 65.87% accuracy and 61.83% F-score for the binary classification (positive vs. negative), with a combined set of syntactic (word-stem), morphological and semantic features.

In a later study, Mourad and Darwish [120] conduct a set of investigations on a collection of nearly 2k Arabic tweets manually annotated for sentiment analysis. The authors utilise a set of syntactic (word-stem), semantic, part-of-speech tagging (POS), stylistics and Twitter-specific features. Training NB and SVM classifiers, they report average scores of 71.9% for accuracy and 70.35% F-score with 10-fold CV setting on the binary task (positive vs. negative). The authors shared their data-set with us. Therefore, we conduct a number of experiments to assess the impact of our expanded feature-sets on the M&D data-set (section 4.2).

A subsequent study by Duwairi et al. [60] on Arabic tweets used a data-set of 1k instances focusing on Jordanian dialect and MSA. The data-set was manually annotated using a crowdsourcing method. The authors experimented with different ML classifiers and performed three-way classification: positive vs. negative vs. neutral. They reported their best performance at an accuracy score of 76.78% using an NB classifier and 5-fold CV setting.

A recent work by Nabil et al. [122] presents an Arabic corpus of 10k tweets manually annotated for SA by three annotators using Amazon Mechanical Turk.

The tweets are collected by querying the top 30 active Egyptian accounts and most trending hashtags in Egypt. Our data-sets (section 3.1 on page 40) are different from Nabil et al. [122]'s data in avoiding bias towards a particular group of users, e.g. active users, and in not targeting a particular dialect. In addition, our data-set is annotated with an extended set of features (see section 3.3 on page 62). Nabil et al. [122] only extract word-based n-grams and for that, they did not report whether they consider a particular lexical representation (e.g. surface or stem forms). For SA prediction, the authors perform 4-way classification: positive vs. negative. vs. neutral vs. mixed. They report the best results with SVM evaluated on a 20% split of the original data-set at an F-score of 0.626% and accuracy of 69.1%.

Conclusion: Studies on SA of Arabic tweets also suffer from a number of shortcomings. For instance, some studies have only targeted a particular dialect, as in [9, 60, 98]. Others have considered only word-based n-gram features, e.g. [13]. In this work, we further expand previous work for SA on Arabic tweets by investigating the impact of: 1) expanded and more variant feature-sets, and 2) experimenting on larger and multi-dialectal training data. In addition, we test our models on an independent test-set, collected at different points in time to explore the performance of our models for a dynamic medium like Twitter. In contrast, Mourad and Darwish [120] and Duwairi et al. [60] only use CV to evaluate their classifiers, while Abdul-Mageed et al. [8] and Nabil et al. [122] use a held-out test-set, which is a sub-set of the original data set used for training. This can be less effective for real-world applications wherein the task is to use trained models for classifying a sample of Twitter feeds over a period of time (section 4.3.4).

Paper	**Data** (size)	**ML Scheme**	**Results**
Abdul-Mageed et al. [7]	newswire (2.8k sentences)	SVM	95.52% acc.
Farra et al. [72]	reviews (44 instances)	SVM	89.3% acc.
Abdul-Mageed et al. [8]	tweets (3k instances)	SVM (held-out)	65.87% acc. and 61.83% F-score
Abbasi et al. [1]	forums (1k instances)	SVM (CV)	93.60% acc.
Mourad and Darwish [120]	tweets (<2k instances)	SVM and NB (CV)	71.9% acc. and 70.35% F-score
Duwairi et al. [60]	tweets (1k Jordanian and MSA)	NB (CV)	76.78% acc.
Nabil et al. [122]	tweets (10k Egyptian)	SVM (held-out)	69.10% acc. and 62.60% F-score

Table 4.1: Summary of previous work on supervised learning SA for Arabic.

.

4.2 Experiments on M&D Data-set

This section describes experiments we conducted on the M&D data-set developed by Mourad and Darwish [120] (page 42), aiming to assess the impact of our new feature-sets. The features we use are summarised in table 3.12 on page 63. Class distribution of M&D data-set is displayed in figure 4.1. We train SVM classifiers using M&D data (ML schemes are described in section 3.5 on page 70). The authors experiment with a set of word-based n-grams, semantic, POS, stylistic, and Twitter-specific features and use an SVM with CV setting. Results of our experiments on this data-set are displayed in table 4.2. The significance of the results are calculated as described in section 3.6.3 on page 75.

Subjectivity classification (polar vs. neutral): The best performance is achieved with the morphological features at 66.25% accuracy. This is 2.65% accuracy improvement compared to the top score originally reported by Mourad and Darwish [120] at 63.6% on this data-set. The addition of the morphological features has significantly improved performance over the stem n-grams baseline. Our morphological feature-set includes POS with 35 tags (see table 3.13 on page 64), as opposed to 5 POS tags used by Mourad and Darwish [120]. We therefore concluded that a rich set of morphological features (e.g. gender, voice, aspect, among others)

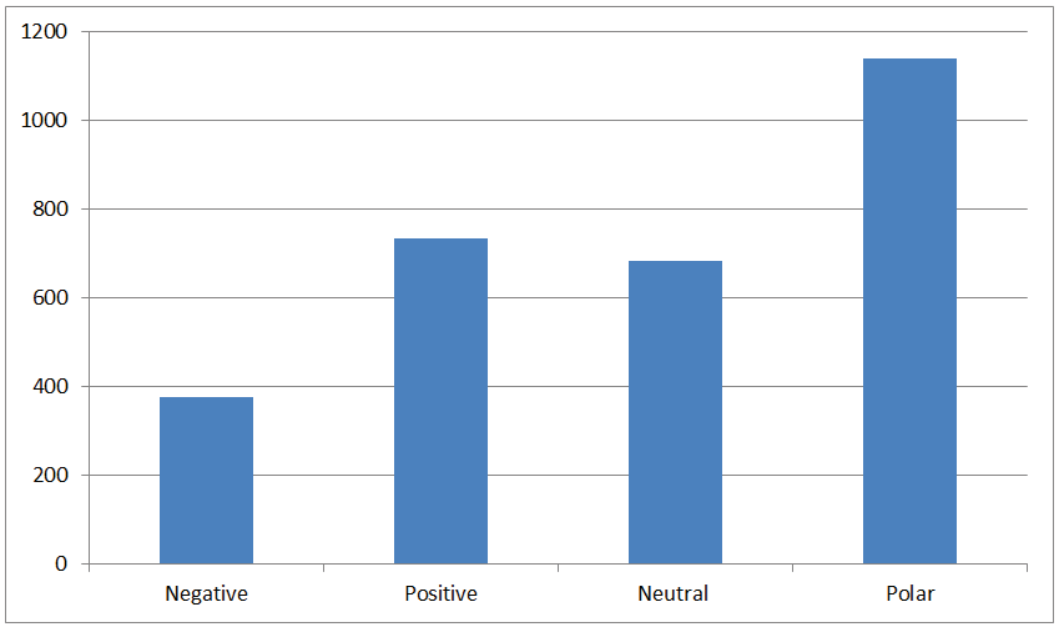


Figure 4.1: Class distribution in M&D data-set.

with an extended POS set is beneficial for Arabic SA.

Sentiment classification (positive vs. negative): The average accuracy score is at 81.32%, which is 9.42% improvement as compared to 71.9% accuracy reported by Mourad and Darwish [120] on this task. The best performance is attained by the semantic features at 82.70% accuracy. For extracting the semantic features, Mourad and Darwish [120] used ArabSenti and a translated version of MPQA, which is similar to our work. However, they did not report on manually correcting/filtering the auto-translated entries of the MPQA in order to maintain its quality. We used a translated and manually filtered version of MPQA that comprises 2.6k entries out of 8k in the original English MPQA (page 64). In addition, they automatically expand the sentiment lexicon, which is likely to introduce more noise than benefit [165]. In our work, we utilised a new dialectal sentiment lexicon to adapt to the use of DAs in social media. We also note that the language-style feature-set has significantly reduced the performance compared to stem baseline. A possible explanation is that this feature-set attempts to capture a correlation between the sentiments being expressed and one or several patterns of informality typically encountered in tweets, e.g. use of repeated letters, ungrammatical use of punctuation, etc. However, de-

M&D Data-set

	Polar vs. Neutral			**Positive vs. Negative**		
	F	Acc.	SD	F	Acc.	SD
Majority baseline (B-mjr)	0.519	65.57	0.17	0.526	66.07	0.4
Stem n-grams $	0.620	65.13	2.81	0.818	82.05	2.64
Stem n-grams + Morph $	0.643	**66.25***	2.54	0.811	81.18	3.99
Stem n-grams + Semantic $	0.620	65.17	2.85	**0.827**	**82.70***	3.56
Stem n-grams + Affec-cues	0.624	65.27	2.87	0.816	81.85	2.93
Stem n-grams + Twt-topic	0.620	65.13	2.82	0.818	82.05	2.62
Stem n-grams + Lang-style $	0.623	63.12*	3.51	0.776	77.61*	4.01
Stem n-grams + Twt-specific $	0.622	65.28	2.78	0.822	82.38	2.92
Comb. of all feat.	**0.65**	66.14*	2.76	0.808	80.78	3.74
Average	0.628	65.19	2.88	0.812	81.32	3.54

Table 4.2: Binary classification on M&D data-set: polar vs. neutral; positive vs. negative. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$). $ denotes that the feature-set or a subset of it has been used by Mourad and Darwish [120].

tecting distinguishable patterns is possibly difficult on such a limited data-set (<2k tweets).

Use of M&D data-set in other studies: A recent study by Salameh et al. [153] on M&D data-set (positive vs. negative) with 10-fold CV and an SVM classifier reported their best score at 74.62%. This is still not competing with our results on this data-set at an average accuracy score of 81.32%. The performance variation can be attributed to the different feature-sets used. Salameh et al. [153] employed word-lemma n-grams and semantic features (leveraging manually and auto-generated sentiment lexica), while our system employs word-stem n-grams along with a wide set of semantic (manually created lexica), a rich set of morphological features, among others.

In sum, our new, extended feature-sets have shown to outperform previous work on M&D data-set for both tasks: subjectivity (polar vs. neutral) and sentiment (positive vs. negative) classification. We did not report on the three-way classification task here because it was not done by Mourad and Darwish [120] (see page 209). In particular, the morphological features have shown to improve accuracy for polar vs. neutral and semantic features have shown to improve accuracy for positive vs. negative.

4.3 Experiments on GS1 Data-set

In the following, we experiment using 10-fold cross-validation on our own gold-standard data-set GS1 (see table 3.8 on page 55). In a second step, the trained models are re-evaluated on our independent test-set to assess their ability to generalise on test instances collected randomly at different points in time (see table 3.9 on page 56). The class distribution of the GS1 data-set is displayed in figure 4.2. As described in section 3.4 (page 69), we conduct a set of experiments considering two different settings: two-level binary classification and single-level three-way.

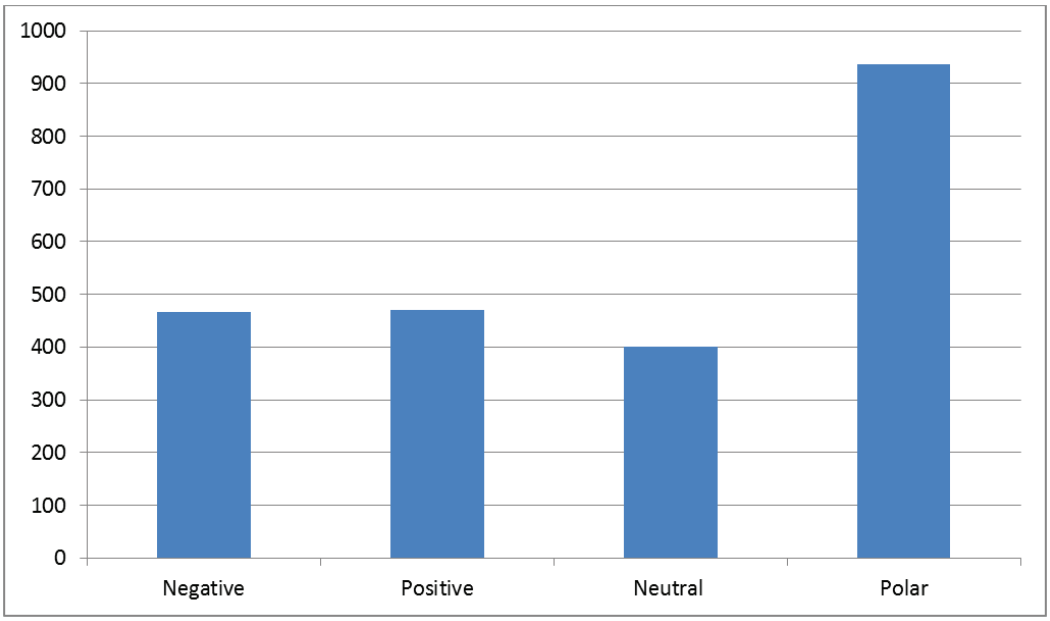


Figure 4.2: Class distribution in GS1 data-set.

4.3.1 Binary classification: Polar vs. Neutral

We first experiment with identifying neutral vs. neutral instances. Results are displayed in table 4.3.

10 Fold Cross-Validation (CV): All the SVMs trained on different feature-sets significantly outperform the majority baseline. The highest performance is achieved when the morphological feature-set is added, which is a significant gain of 9.4% accuracy over using stem n-grams only. The average accuracy across all feature-sets is at 95.49%.

Polar vs. Neutral

	10 Fold CV			Ind. Test-set	
	F	Acc.	SD	F	Acc.
Majority baseline (B-mjr)	0.578	70.08	0.1	0.471	61.70
Stem n-grams	0.905	91.01	2.24	0.557	65.26
Stem n-grams + Morph	**0.999**	99.85*	0.35	0.596	65.58*
Stem n-grams + Semantic	0.906	91.11	2.25	0.562	65.46*
Stem n-grams + Affec-cues	0.906	91.15	2.2	0.565	65.52*
Stem n-grams + Twt-topic	0.905	91.01	2.24	0.576	**65.97***
Stem n-grams + Lang-style	0.990	99.80*	0.35	0.536	62.35*
Stem n-grams + Twt-specific	0.998	99.92*	0.23	**0.627**	63.82*
Comb. of all feat.	0.998	**99.93***	0.23	0.594	63.14*
Average	0.952	95.49	1.86	0.577	64.26

Table 4.3: Binary classification on GS1: polar vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

Independent test-set: The purpose of this experiment is to evaluate the ability of our models to perform polarity classification for a time-changing platform like Twitter. The stem n-grams set a strong baseline at 65.26% accuracy. The best individual contribution is recorded with the Twitter-topic feature with 0.71% accuracy improvement over the stem baseline (table 4.4). Table 4.4 shows that the Twitter-topic feature-set attained the lowest classification error at 0.3391 with an effect size of 0.85. Overall, we can observe a significant performance drop of 31.23% accuracy on average between CV and the results on independent test-set. This indicates that, despite the promising results with CV at an average accuracy of 95.49%, the classifiers do not generalise well to unseen topics.

4.3.2 Binary classification: Positive vs. Negative

This set of experiments distinguishes between positive and negative sentiments. Table 4.5 summarises the results.

10 Fold Cross-Validation (CV): All classifiers significantly outperform the majority baseline. The best performance is attained with stem+morph at 90.87% accuracy (+16.77%) followed by stem+language-style features at 88.82% accuracy (+14.72%) over the stem baseline. The average accuracy across all feature-sets is at 80.24%.

	Polar vs. Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	2847.449 (0.000)	0.897	0.3470
Stem n-grams + Morph	1135.109 (0.000)	0.566	0.3742
Stem n-grams + Semantic	2688.258 (0.000)	0.871	0.3420
Stem n-grams + Affec-cues	2648.772 (0.000)	0.871	0.3434
Stem n-grams + Twt-topic	2577.841 (0.000)	0.853	0.3391
Stem n-grams + Lang-style	2282.918 (0.000)	0.803	0.3801
Stem n-grams + Twt-specific	592.628 (0.000)	0.409	0.3617
Comb. of all feat.	1007.503 (0.000)	0.533	0.3663

Table 4.4: Comparison of performance using different feature-sets of GS1 data-set on the independent test-set.

Independent test-set: All classifiers outperformed the stem baseline. The only exception is with the affective-cues features that resulted in a marginal drop of 0.5% lower than the stem baseline. The best individual contribution here is recorded with stem+morph feature-set at 68.99% accuracy that has significantly outperformed the stem-baseline. The top performance is attained when all features are combined at an accuracy score of 69.68%, which is 11.09% significant improvement over the stem baseline (table 4.6). Again, testing on the independent test-set has resulted in an average drop of 17.03% accuracy across all feature-sets as compared to CV.

4.3.3 Three-way classification: Positive vs. Negative vs. neutral

We now experiment with single level three-way classification for positive vs. negative vs. neutral. Table 4.7 summarises the results.

10 Fold Cross-Validation (CV): All classifiers have significantly outperformed the majority baseline. As for the individual blocks of features, the addition of morphological features has resulted in the biggest improvement, attaining an accuracy

Positive vs. Negative					
	10 Fold CV			Ind. Test-set	
	F	Acc.	SD	F	Acc.
Majority baseline (B-mjr)	0.335	50.16	0.25	0.531	66.51
Stem n-grams	0.736	74.1	3.71	0.586	58.59
Stem n-grams + Morph	**0.909**	**90.87***	2.59	0.694	68.99*
Stem n-grams + Semantic	0.752	75.45*	3.81	0.690	68.16*
Stem n-grams + Affec-cues	0.732	73.75*	3.8	0.581	58.04
Stem n-grams + Twt-topic	0.736	74.11	3.7	0.595	59.32
Stem n-grams + Lang-style	0.89	88.82*	2.5	0.635	63.31*
Stem n-grams + Twt-specific	0.736	74.08	3.75	0.599	59.60
Comb. of all feat.	0.908	90.77*	2.41	**0.702**	**69.68***
Average	0.80	80.24	3.29	0.635	63.21

Table 4.5: Binary classification on GS1: positive vs. negative. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive vs. Negative		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	617.63 (0.000)	0.531	0.4141
Stem n-grams + Morph	18.56 (0.000)	0.090	0.3101
Stem n-grams + Semantic	366.26 (0.000)	0.409	0.3142
Stem n-grams + Affec-cues	1254.50 (0.000)	0.757	0.4438
Stem n-grams + Twt-topic	1093.01 (0.000)	0.707	0.4182
Stem n-grams + Lang-style	65.89 (0.000)	0.172	0.4008
Stem n-grams + Twt-specific	934.30 (0.000)	0.654	0.4076
Comb. of all feat.	73.46 (0.000)	0.182	0.3055

Table 4.6: Comparison of performance using different feature-sets of GS1 data-set on the independent test-set.

Positive vs. Negative vs. Neutral

	10 Fold CV			**Ind. Test-set**	
	F	Acc.	SD	F	Acc.
Majority baseline (B-mjr)	0.183	35.15	0.12	0.239	41.04
Stem n-grams	0.745	74.22	3.42	0.425	43.78
Stem n-grams + Morph	**0.937**	**93.68***	1.91	0.484	48.59*
Stem n-grams + Semantic	0.754	74.99*	3.36	0.467	48.84*
Stem n-grams + Affec-cues	0.744	74.16	3.57	0.428	43.84
Stem n-grams + Twt-topic	0.745	74.16	3.44	0.438	44.72*
Stem n-grams + Lang-style	0.926	92.57*	2.18	0.380	40.45*
Stem n-grams + Twt-specific	0.813	81.65*	2.94	0.451	44.72*
Comb. of all feat.	0.935	93.51*	2.04	**0.496**	**49.75***
Average	0.825	82.37	2.85	0.446	45.58

Table 4.7: Three-way classification on GS1: positive vs. negative vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive vs. Negative vs. Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	3438 (0.000)	0.985	0.5638
Stem n-grams + Morph	11.156 (0.004)	0.053	0.5028
Stem n-grams + Semantic	2128 (0.000)	0.775	0.5033
Stem n-grams + Affec-cues	3354 (0.000)	0.973	0.5633
Stem n-grams + Twt-topic	3120 (0.000)	0.939	0.5537
Stem n-grams + Lang-style	742.9 (0.000)	0.458	0.5986
Stem n-grams + Twt-specific	1424 (0.000)	0.634	0.5570
Comb. of all feat.	93.72 (0.000)	0.162	0.4951

Table 4.8: Comparison of performance using different feature-sets of GS1 data-set on the independent test-set.

score of up to 93.68%. This is 19.46% accuracy improvement over the stem baseline. The average accuracy score across all feature-sets is at 82.37%.

Independent test-set: The combination of all feature-sets attained the best scores for re-evaluating the models on the independent test-set at 49.75% accuracy, significantly outperforming a stem baseline at 43.78%. Table 4.8 displays that the combination of all features attained the lowest classification error at 0.495 with a small effect size of 0.162. Compared to CV, there is an average performance drop by 36.79% accuracy.

4.3.4 Summary of GS1 Results

In summary, the GS1 experiments revealed:

- The stem-based n-grams features set a strong baseline as confirmed by previous work, e.g. [129, 7, 6].

- The experimental design helps assessing the individual contributions of different blocks of feature-sets. A combination of all features does not necessarily yield to the top performance. This is possibly because some feature-sets might hurt the performance, e.g. language-style in polarity and three-way classification, while other feature-sets can have no/negligible effects, e.g. affective-cues in three-way classification.

- With Arabic as a morphologically rich language (section 2.3.3 on page 24), amongst our best performing feature-sets are morphological features. This confirms findings by Abdul-Mageed et al. [8], which observe an improvement in performance when adding more morphological features. Despite noise introduced by the morphological analyser MADAMIRA (see page 62), which is designed for MSA only, the extracted rich set of morphological features remain useful for SA on Arabic tweets.[1] A possible explanation is that "variations of some of the morphological features (e.g., existence of a gender, person feature) may correlate more frequently with positive or negative sentiment" [8]. A closer look at morph features of our GS data reveals that there is more masculine in negative tweets than positive ones, and there is more use of singular form in positive tweets than negative ones.

- The two-level binary classification model leads to better results, with an average accuracy of 63.74%, as compared to 45.58% achieved with the one-level three-way classification model. This is expected since multi-class classification is likely to be a more difficult task than binary classification.

[1] For the entire GS data-set, the tweets with dialectal expressions represent 24%. Within subjective/polar tweets, 44.03% of the tweets are dialectal and majority of the dialectal instances are negative.

- Unlike previous work, we re-evaluate our trained models on an independent, larger and more diverse test-set. We show that, despite very promising CV results, our models do not generalise well to data-sets collected at a later point in time, causing a performance drop of 24.13% accuracy on average (table 4.24).

- The performance drop is likely to be caused by time-dependent topic-shifts issues in the Twitter stream and the prominent role of word n-gram features in our models [128, 165]. Since Twitter experiences topic-shifts over time, the vocabulary, especially the content words, are likely to change as well [61]. Investigating this hypothesis, we find that the word frequency distribution differs amongst the training/test data-sets: the overall overlap of unique tokens is only 12.21%. We will address this issue by using a larger gold-standard training data-set (section 4.4) and by using semi-supervised approaches to automatically obtain larger training data (chapter 5).

4.4 Experiments on GS2 Data-set

In this section, we investigate the effect of the same experimental settings and feature-sets used in the previous section but on a larger data-set, which is GS2 (see page 55). GS2 is composed of 6.8k tweets, which is 3 times larger than GS1. The purpose of this set of experiments is assess the impact of using GS2 as a larger gold-standard training data on alleviating the performance drop encountered on the independent test-set in our previous set of experiments (section 4.3.4). In addition, we evaluate the performance of models trained on MSA instances only (B-MSA, see page 72) against models trained on MSA+DA instances to explore the impact of DA presence on the overall performance of the sentiment classifiers (see figure 4.3). To automatically detect MSA vs. DA tweets, we use AIDA, a publicly available tool that distinguishes MSA vs. DA instances (page 22) [66]. Because GS2 is 3 times larger in size than GS1, we decided to run this investigation (MSA vs. MSA+DA) only on GS2, as it is likely to better show the effect of MSA/DA distribution.

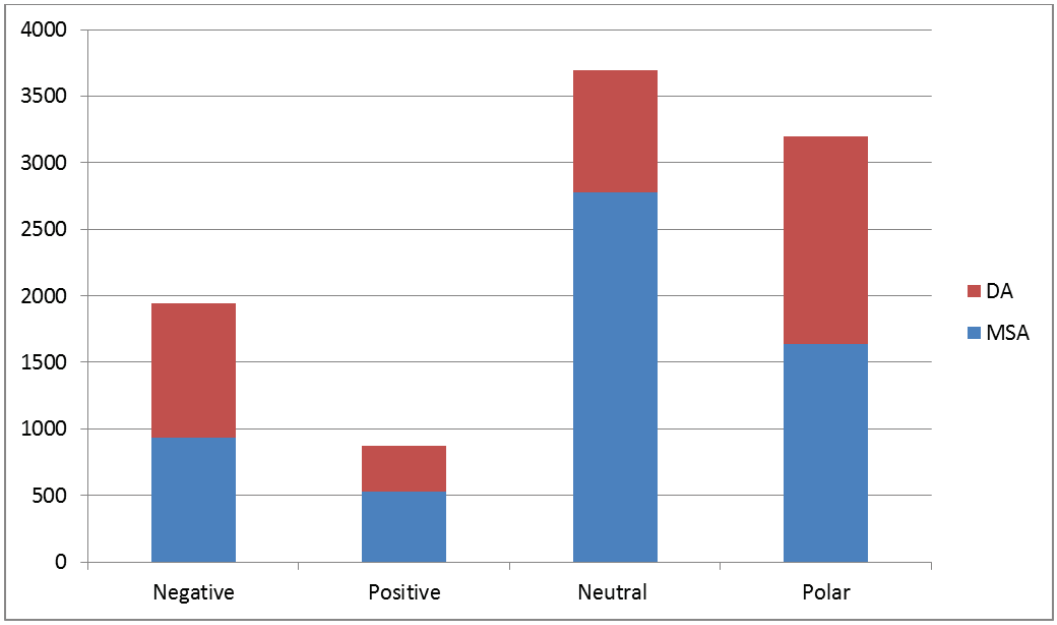


Figure 4.3: Distribution of MSA/DA instances within each class of GS2 data-set.

4.4.1 Binary classification: Polar vs. Neutral

The results for discriminating polar vs. neutral instances are displayed in table 4.9.

10 Fold Cross-Validation (CV): All classifiers significantly outperform a majority baseline. The combination of all features have achieved the best performance with an improvement of 3.96% over a stem-baseline of 72.72% accuracy.

For MSA baseline, the addition of 2,478 dialectal instances (representing 35.94% of data-set) has resulted in an average improvement of 1.39% in accuracy over a baseline model trained only on MSA tweets. It can be inferred that the biggest contribution in the overall performance of MSA+DA is caused by the MSA instances. That is, with B-MSA instances (representing only 62% of the training set), the classifiers are able to attain an average accuracy of up to 72.0% as compared to 73.39% achieved with the entire training-set of MSA+DA.

Independent test-set: The average accuracy attained on this task is at 70.59% accuracy. Interestingly, the performance drop between CV and independent test-set settings is only an average of 2.86% as compared to 31.23% for GS1 on the same task, which shows a more consistent/stable performance as a result of exploiting a larger training set, i.e. models are better able to generalise. The best individual contribution here is achieved when the Twitter-topic feature-set is used (+0.66% accuracy over the stem baseline), suggesting a utility for non-word token based features in this context. Table 4.10 shows that the Twitter-topic attained the lowest classification error at 0.300 with a medium effect size of 0.33.

The MSA baseline (B-MSA) attain an average accuracy of up to 57.61% on this task (the MSA/DA distributions of GS2 data-set is shown in figure 4.3). The MSA/DA distribution of the test-set is displayed in figure 4.4). The addition of of tweets identified as dialectal has resulted in 12.98% improvement in average accuracy recorded. A similar behaviour was also reported by Mourad and Darwish [120]. It appears that, despite their noise, the presence of DA instances in training data boost the models performance.

Polar vs. Neutral

	10 Fold CV						**Ind. Test-set**			
	B-MSA			*MSA+DA*			*B-MSA*		*MSA+DA*	
	F	Acc.	SD	F	Acc.	SD	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.399	55.73	0.05	53.63	0.374	0.06	0.471	61.70	0.471	61.70
Stem n-grams	0.714	71.43	1.56	0.727	72.72	1.68	0.492	53.51	0.718	71.76
Stem n-grams + Morph	0.733	73.29*	1.69	0.747	74.73*	1.78	0.667	66.28*	0.656	68.70*
Stem n-grams + Semantic	0.711	71.14	1.62	0.726	72.66*	1.63	0.485	53.0	0.718	71.85
Stem n-grams + Affec-cues	0.713	71.41	1.53	0.727	72.69	1.62	0.497	53.82	0.718	71.78
Stem n-grams + Twt-topic	0.719	71.90	1.68	0.73	73.05*	1.62	0.457	51.25*	**0.724**	**72.42***
Stem n-grams + Lang-style	0.702	70.49*	1.86	0.72	72.18	4.0	0.585	59.53*	0.682	68.21*
Stem n-grams + Twt-specific	0.726	72.61*	1.61	0.734	73.40*	1.71	0.583	59.39*	0.702	70.95*
Comb. of all feat.	**0.737**	**73.68***	1.66	**0.757**	**75.67***	1.8	0.645	64.08*	0.677	69.05*
Average	0.719	72.0	1.65	0.734	73.39	2.13	0.552	57.61	0.699	70.59

Table 4.9: Binary classification on GS2: polar vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Polar vs. Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	533.725 (0.000)	0.388	0.3216
Stem n-grams + Morph	410.733 (0.000)	0.340	0.3278
Stem n-grams + Semantic	543.356 (0.000)	0.391	0.3159
Stem n-grams + Affec-cues	541.745 (0.000)	0.391	0.3168
Stem n-grams + Twt-topic	398.214 (0.000)	0.331	0.3006
Stem n-grams + Lang-style	681.803 (0.000)	0.438	0.3360
Stem n-grams + Twt-specific	299.023 (0.000)	0.290	0.3058
Comb. of all feat.	93.105 (0.000)	0.162	0.3277

Table 4.10: Comparison of performance using different feature-sets of GS2 data-set on the independent test-set.

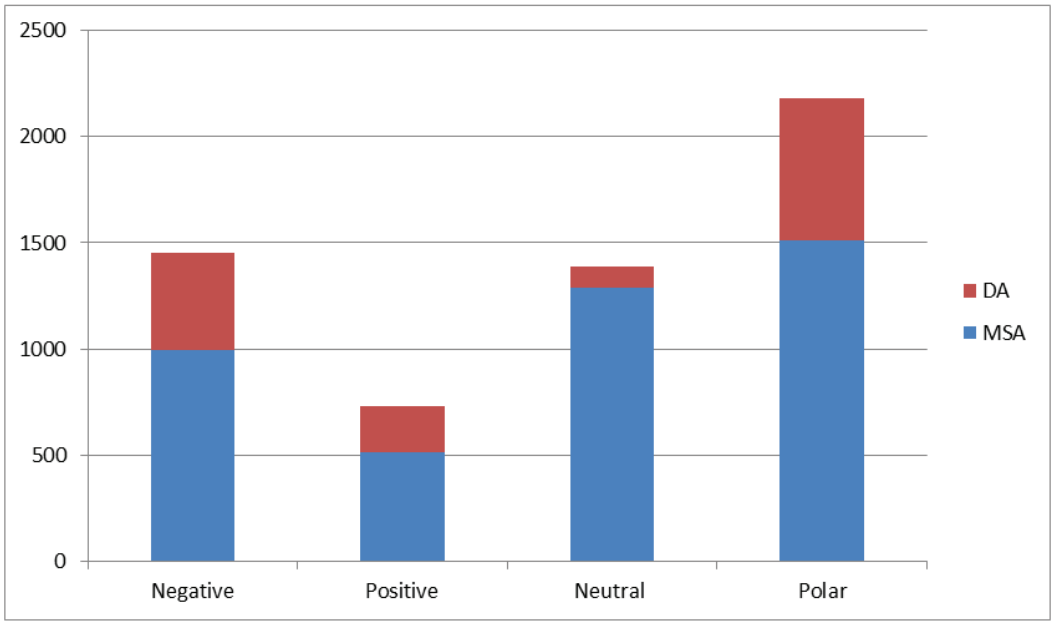


Figure 4.4: Distribution of MSA/DA instances within each class of the independent test-set.

4.4.2 Binary classification: Positive vs. Negative

The results for discriminating positive vs. negative instances are displayed in table 4.11 and values of effect size and classification error are shown in table 4.12.

10 Fold Cross-Validation (CV): All classifiers significantly outperform majority baseline. The best performance is attained with stem+morph at 89.40%, which is 1.4% accuracy improvement over the stem baseline.

The average accuracy score with GS2 under CV setting is at 87.72%, which is 7.48% better than average accuracy recorded on GS1 under the same setting. Note that we directly compare the performance of GS1 and GS2 classifiers by evaluating them against the independent test-set (page 99).

For MSA baseline (B-MSA), with only MSA instances, the models attained an average accuracy of 77.75%. Adding a set of 1,353 dialectal instances (representing a total of 48.03% of the training-set) has resulted in an improvement of 9.97% over MSA baseline. It is worth noting that the addition of the dialectal tweets has resulted in only 1.39% accuracy improvement on subjectivity classification (section 4.4.1). A possible explanation is that DA proportion in this task (positive vs. negative) is larger, representing 48.03% of data-set, compared to 35.94% with polar vs. neutral

task.

Use of GS2 data-set in other studies: It is interesting to mention that a recent study by Salameh et al. [153] has used our GS2 data-set and reported the best accuracy score at 85.23% with 10-fold CV (positive vs. negative). They use an SVM classifier and exploit word-lemma n-grams and semantic features. This is still not outperforming our best results on this data-set at 89.40% using word-stem n-grams and a rich set of automatically extracted morphological features (table 4.11).

Independent Test-set: The best individual contribution here is reached when using the affective-cues feature-set at 78.22% (+0.48% over the stem baseline), indicating their usefulness for this task of binary sentiment classification. Table 4.12 indicates that the affective-cues feature-set attained the lowest classification error in this set of experiments at 0.227 with a medium effect size of 0.38. Overall, the performance on the independent test-set attained an average accuracy at 75.32%, which is 12% better than that achieved by models trained on the smaller GS1 at 63.21% on the same task. As previously mentioned, GS2 is >3 times larger in size than GS2. This confirms the utility of exploiting a larger training data in improving the models' ability to generalise.

With respect to the MSA baseline (B-MSA), the model trained only on tweets identified as MSA has reached an average accuracy score of 72.07%. The addition of the dialectal tweets has resulted in an improvement of up to 3.25% accuracy. It is interesting to note that DA instances have been shown useful for both subjectivity (polar vs. neutral) and sentiment (positive vs. negative) classification. However, the results suggest that the presence of DA instances can be more beneficial for a model discriminating polar (more likely to be dialectal) vs. neutral (more likely to be MSA) (section 4.4.1 on page 96).

Positive vs. Negative

	10 Fold CV						**Ind. Test-set**			
	B-MSA			*MSA+DA*			*B-MSA*		*MSA+DA*	
	F	Acc.	SD	F	Acc.	SD	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.463	61.05	0.07	0.362	52.56	0.14	0.531	66.51	0.531	66.51
Stem n-grams	0.780	78.34	2.59	0.88	88.01	1.96	0.702	74.85	0.763	77.74
Stem n-grams + Morph	**0.796**	**79.74***	2.42	**0.894**	**89.40***	1.77	0.656	70.55*	0.715	74.66*
Stem n-grams + Semantic	0.783	78.58	2.75	0.881	88.09	2.09	0.689	74.26	0.762	77.84
Stem n-grams + Affec-cues	0.777	77.92	2.61	0.877	87.69	1.90	0.706	74.99	**0.769**	**78.22**
Stem n-grams + Twt-topic	0.779	77.89*	2.60	0.880	88.03	2.05	0.591	68.76*	0.697	74.65*
Stem n-grams + Lang-style	0.730	73.22*	2.87	0.847	84.70*	2.35	0.681	71.60*	0.70	70.58*
Stem n-grams + Twt-specific	0.777	77.78	2.56	0.881	88.12	1.99	0.684	73.34	0.768	78.13
Comb. of all feat.	0.784	78.54	2.38	0.879	87.89	2.03	0.588	68.21*	0.659	70.72*
Average	0.776	77.75	2.59	0.877	87.72	2.02	0.662	72.07	0.729	75.32

Table 4.11: Binary classification on GS2: positive vs. negative. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive vs. Negative		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	328.262 (0.000)	0.387	0.2327
Stem n-grams + Morph	222.833 (0.000)	0.319	0.2624
Stem n-grams + Semantic	375.058 (0.000)	0.414	0.2359
Stem n-grams + Affec-cues	318.286 (0.000)	0.382	0.2272
Stem n-grams + Twt-topic	625.370 (0.000)	0.535	0.2817
Stem n-grams + Lang-style	159.156 (0.000)	0.270	0.2858
Stem n-grams + Twt-specific	303.611 (0.000)	0.373	0.2423
Comb. of all feat.	490.209 (0.000)	0.474	0.2853

Table 4.12: Comparison of performance using different feature-sets of GS2 data-set on the independent test-set.

4.4.3 Three-way classification: Positive vs. Negative vs. Neutral

The results for classifying positive vs. negative vs. neutral instances are displayed in table 4.13.

10 Fold Cross-Validation (CV): All classifiers significantly outperform a majority baseline. The best score is attained with the combination of all feature-sets at 76.97%. This is 2.54% accuracy improvement over the stem baseline. The best individual performance for a feature-set is achieved with the morphological features, with a significant gain of 1.5% over the stem n-grams baseline. The average performance across all feature-sets is at 76.97% accuracy.

Independent test-set: The best individual contribution is attained by the addition of Twitter-specific feature-set at 60.74% accuracy, which is 1.10% significant improvement over the stem baseline. Table 4.14 shows that the Twitter-specific feature-set attained the lowest classification error at 0.397 with a small effect size of 0.12. This suggests a utility for this feature-set to enhance performance independent of topic/temporal-related issues. For instance, looking at the data-set reveals that: the number of tweets with *is-Favourite:true* tends to be balanced between classes, while *is-Retweet:true* tends to be more frequent with positive tweets and *has-hashtag:true* tends to appear more frequently with negative tweets. The average performance on the independent test-set across all feature-sets is at 59.29%, which is significantly better that that achieved by GS1 on the same task at 45.58%. This shows that models trained on larger data-sets are better able to generalise.

With respect to the MSA baseline (B-MSA), the addition of tweets identified as dialectal (representing 23.01% of data-set) has resulted in an improvement of 6.01% in average accuracy score. This is in line with the results on binary tasks (subjectivity and sentiment) classification, suggesting a consistent positive impact for the existence of DA instances for SA classifiers.

Positive vs. Negative Vs. Neutral

	10 Fold CV						**Ind. Test-set**			
	B-MSA			*MSA+DA*			*B-MSA*		*MSA+DA*	
	F	Acc.	SD	F	Acc.	SD	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.437	58.92	0.05	0.334	50.03	0.07	0.239	41.04	0.239	41.04
Stem n-grams	0.655	67.48	1.73	0.74	74.43	1.65	0.472	52.04	0.577	59.64
Stem n-grams + Morph	0.685	69.31*	1.69	0.758	75.93*	1.52	0.545	55.21*	0.584	58.93*
Stem n-grams + Semantic	0.654	67.32	1.74	0.739	74.37	1.70	0.466	51.61	0.576	59.61
Stem n-grams + Affec-cues	0.659	67.71	1.70	0.743	74.67	1.66	0.480	52.61	0.583	60.18
Stem n-grams + Twt-topic	0.664	67.99*	1.80	0.747	75.06*	1.60	0.481	52.63	0.588	60.63*
Stem n-grams + Lang-style	0.65	66.08*	1.90	0.731	73.51	1.58	0.503	53.25*	0.546	56.93*
Stem n-grams + Twt-specific	0.665	67.81	1.78	0.746	74.85*	1.63	0.528	55.66*	**0.594**	**60.74***
Comb. of all feat.	**0.687**	**69.55***	1.63	**0.768**	**76.97***	1.81	0.511	53.28*	0.557	57.63*
Average	0.665	67.91	1.74	0.747	74.98	1.64	0.498	53.28	0.576	59.29

Table 4.13: Three-way classification on GS2: positive vs. negative vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive vs. Negative VS Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	292.677 (0.000)	0.287	0.4180
Stem n-grams + Morph	439.272 (0.000)	0.309	0.4200
Stem n-grams + Semantic	297.363 (0.000)	0.289	0.4191
Stem n-grams + Affec-cues	261.187 (0.000)	0.271	0.4075
Stem n-grams + Twt-topic	183.215 (0.000)	0.227	0.4101
Stem n-grams + Lang-style	92.198 (0.000)	0.161	0.4262
Stem n-grams + Twt-specific	51.282 (0.000)	0.120	0.3976
Comb. of all feat.	340.217 (0.000)	0.309	0.4296

Table 4.14: Comparison of performance using different feature-sets of GS2 data-set on the independent test-set.

4.4.4 Summary of GS2 Results

In summary, the GS2 experiments revealed:

- Training models using a 3 times larger data-set has resulted in closing/reducing the gap in performance between CV and independent test-set evaluation settings from 30.46% on GS1 to only 11.66% on GS2 (table 4.24 on page 115).

- On the independent test-set, the average accuracy for binary classification is 9.22% better using GS2 than that recorded with GS1. As for three-way classification, the average accuracy improved by 13.71% using GS2.

- The B-MSA baseline reveals a general positive impact for the addition of DA instances to training data. Specifically, the results show that adding a 1/3 of the training-set comprising only DA instances has resulted in an improvement of 5.68% accuracy for binary classification and 7.07% accuracy for three-way classification. This suggests that in spite of noise potentially introduced with dialectal instances, their presence remains useful as they allow models to be exposed to genre-specific features including dialectal expressions. A similar behaviour is reported by Zbib et al. [185] when performing an MT-based task on Arabic web text.

- Our system is able to outperform the results reported in a recent study by Salameh et al. [153] who have used our GS2 data-set and reported their best accuracy score at 85.23% with 10-fold CV (positive vs. negative). This is still not outperforming our best results on this data-set at 89.40% with the same experimental setting, proving superiority for the features employed by our system.

4.5 Experiments on GS1+GS2 Data-set

Finally, this section assesses the impact of combining data-sets used in previous sections. In particular, we experiment with a merged data-set comprising instances of GS1 and GS2, resulting in 9k instances (table 3.8 on page 55). Figure 4.5 shows the class distribution in this combined data-set. This is motivated by the promising results attained when more training data was used (section 4.4.4) [31]. Banko and Brill [31] showed that using larger training data will lead to an improvement in text classification tasks. In addition, we examine the per-class performance to observe performance variation across different classes. The reason is that we anticipate this combined data (GS1+GS2) to yield a better performance compared to GS1 and GS2 individually. Thus, we aim to compare the performance of the sentiment classifiers trained on GS1+GS2 against the results of state-of-the-art SA systems of SemEval'15 on English tweets [145]. For this, we follow their procedure of assessing per-class F-scores (further discussion in section 4.6).

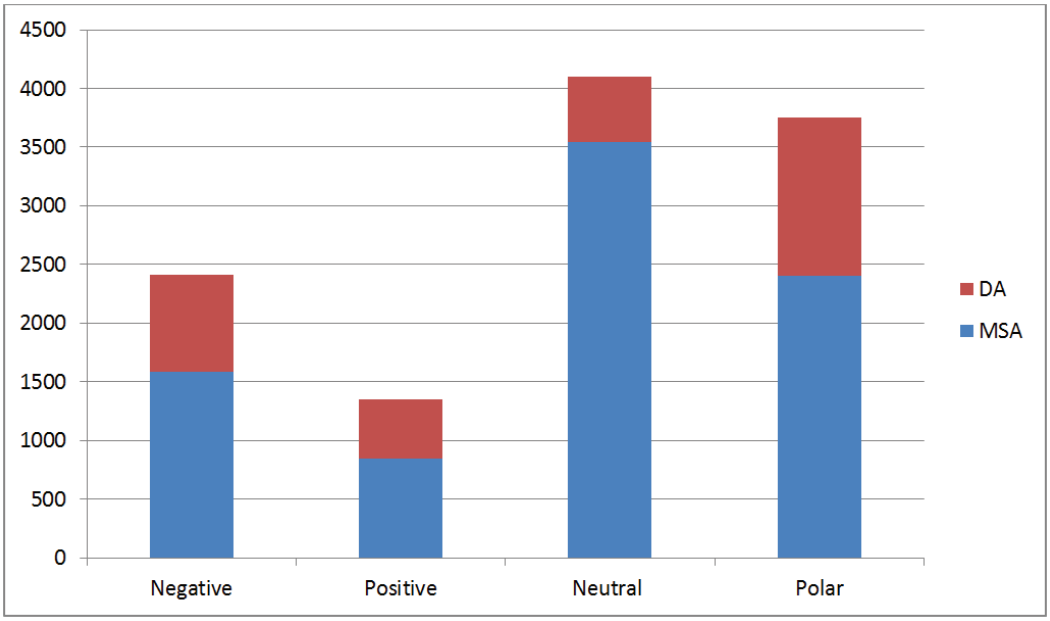


Figure 4.5: Distribution of MSA/DA instances within each class of GS1+GS2 data-set.

4.5.1 Binary classification: Polar vs. Neutral

This section presents the results for classifying polar vs. neutral instances in the combined GS1+GS2 data-set. Results are shown in table 4.15.

10 Fold Cross-Validation (CV): All classifiers significantly outperform a majority baseline. The best performance is attained by combining all feature-sets at 79.68% accuracy, which significantly outperform a baseline of only stem n-grams with an improvement of 3.54%. Generally, all the feature-sets have shown an improvement on this task. This suggests that as the size of data-set increases, the more possibilities can occur for feature-sets to reflect clearer patterns that correlate with a certain sentiment class [31].

Independent test-set: The best performance is achieved by the stem n-grams at an accuracy score of 73.99%, which is better than the score achieved by the combination of all features at 69.62% (table 4.16).

The average accuracy score is at 72.29%, which is 8.03% better than the score recorded with GS1 and 1.7% better than GS2, indicating a positive effect for combining GS1+GS2 on this task.

The per-class F-scores indicate a better performance when detecting the polar class with an average of 0.79 as compared to 0.592 for detecting the neutral class. This is surprising, especially because neutral is the majority class in this task, which is assumed to allow the classifiers to be exposed to more neutral instances, and hence learn more distinguishable aspects of this class. A possible explanation is the presence of good/bad news within the neutral class that might pose a challenge/confusion to the classifiers. In this context, Abdul-Mageed and Diab [4] argue that the inclusion of good/bad news as neutral instances imposes an additional difficulty on the SA classifiers. Similarly, Mourad and Darwish [120] state that news - especially in Arabic - can be reported in an "overly critical" way (see examples in table 4.17 from our data). In this context, Wilson et al. [174] highlight the negative impact of the presence of the positive or negative expressions in a neutral context on the performance of a sentiment classifier. In addition, it is interesting to see that

Polar vs. Neutral

	10 Fold CV			Ind. Test-set			
	F	Acc.	SD	F $_{polar}$	F $_{neut.}$	F $_{(polar,neut.)}$	Acc.
Majority baseline (B-mjr)	0.358	52.18	0.06	0.763	0.000	0.471	61.70
Stem n-grams	0.761	76.14	1.35	0.798	**0.634**	**0.735**	**73.99**
Stem n-grams + Morph	0.783	78.26*	1.21	0.782	0.523	0.683	70.10*
Stem n-grams + Semantic	0.762	76.19	1.43	0.797	0.623	0.730	73.57
Stem n-grams + Affec-cues	0.762	76.17	1.27	0.796	0.631	0.733	73.77
Stem n-grams + Twt-topic	0.764	76.39	1.33	**0.806**	0.592	0.724	73.74
Stem n-grams + Lang-style	0.767	76.82*	1.64	0.786	0.608	0.718	72.36
Stem n-grams + Twt-specific	0.776	77.64*	1.20	0.775	0.598	0.707	71.17
Comb. of all feat.	**0.797**	**79.68***	1.34	0.776	0.527	0.681	69.62*
Average	0.772	77.16	1.35	0.790	0.592	0.714	72.29

Table 4.15: Binary classification on GS1+GS2: polar vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Polar VS Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	530.403 (0.000)	0.387	0.2603
Stem n-grams + Morph	700.313 (0.000)	0.445	0.2987
Stem n-grams + Semantic	826.681 (0.000)	0.483	0.2603
Stem n-grams + Affec-cues	760.330 (0.000)	0.463	0.2689
Stem n-grams + Twt-topic	1363.00 (0.000)	0.620	0.2696
Stem n-grams + Lang-style	217.593 (0.000)	0.247	0.2823
Stem n-grams + Twt-specific	518.074 (0.000)	0.382	0.2795
Comb. of all feat.	796.005 (0.000)	0.474	0.3055

Table 4.16: Comparison of performance using different feature-sets of GS1+GS2 data-set on the independent test-set.

tweet-topic feature-set has a positive effect in discriminating the polar instances. A closer look at the data-set reveals that topics like social/religious issues are predominantly polar/subjective, while tweets under economic issues are predominantly neutral.

4.5.2 Binary classification: Positive vs. Negative

Results for experiments on this binary task are displayed in table 4.20.

10 Fold Cross-Validation (CV): All classifiers significantly outperform a majority baseline. The addition of morphological features has resulted in a significant

1	محَادثَات شَاءيكه بَانتظَار اردوغَان في بروكسل *Prickly conversations awaiting Erdoğan in Brussels.*
2	اقل من ١ % من السعوديّين متعَاطفون مع دَاعش *Less than 1% of Saudis sympathetic with ISIS.*
3	الحرب الطَاءيفيه تنعش تجَاره السلَاح في العرَاق *Arms trade is reviving/flourishing in Iraq because of sectarian war.*
4	ريَال مدريد يقهر برشلونه ويخطف النصر *Real Madrid conquers Barcelona and snatches a victory.*
5	الَاتحَاد الَاوروبي يعرب عن صدمته لمقتل مءيَات المتظَاهرين في مصر *European Union expresses its shock for killing hundreds of protesters in Egypt.*

Table 4.17: Examples of news tweets.

gain of 3.55% accuracy, followed by the combination of all feature-sets with a significant gain of 2.77%. The average accuracy across all feature-sets is at 80.90%.

Independent test-set: The best performances here are recorded with the morphological and Twitter-specific feature-sets, with 1.74% and 1.37% improvements over the stem baseline respectively. Table 4.21 indicates that the morphological features attained the lowest classification error rate at 0.220 with a small effect size at 0.11. The average accuracy performance is 0.59% better than GS2 and 12.7% better than GS1 on this task.

Interestingly, the per-class F-scores indicate that discriminating positive class (avg. F-positive = 0.645) is a harder task than negative class (avg. F-negative = 0.820). Abdul-Mageed et al. [6] noted a similar behaviour with Arabic tweets as they reported F-positive at 0.419 and F-negative at 0.718. This is surprising because, despite being the minority class in our experiment, positive sentiment-bearing words appear to be prevalent in a compiled list of the most predictive/informative features that are expected to give the best clues for sentiment classifiers (see table 4.18). This frequent use of positive words is in line with findings of a recent study by Dodds et al. [58][2] that reveals a universal tendency to use positive words more often

[2]Dodds et al. [58] conduct their study on 24 corpora in 10 languages (including Arabic) from several resources including Twitter and New York Times, among others.

and in a wider range of forms than that used to convey negative emotions across languages. Despite the wide use of positive words, our sentiment classifiers perform less effectively on positive class. To investigate this we look at samples of the data-set closely, which reveal a notable use of positive words in a negative context, either sarcastically or as a means of stressing/amplifying the intended feeling (see examples 6-7 in table 4.19).

	GS1+GS2		
ID	**Arabic**	**English**	χ^2
1	جميل	beautiful	65.2234
2	خير	well-being	52.208
3	مبروك	congratulations	48.653
4	حلو	nice	41.6515
5	دنيَا	world	38.2499
6	رب	Lord	38.2499
7	فرح	happiness	35.0168
8	رَاءيع	gorgeous	32.3793
9	ابتسَامه	smile	30.5824
10	اخو	brother	30.432
11	ارهَاب	terrorism	28.9001
12	شكرَا	thanks	24.616
13	حب	love	23.3379
14	اعلَام	media	22.0386
15	نظَام	system/regime	18.4238

Table 4.18: The most predictive word uni-grams (for positive vs. negative) in the GS1+GS2 data-set as evaluated by Chi-Squared.

6	مبروك خسرت كرَامتك ***Congratulations**, you have lost your dignity.*
7	اسرَاءيّل اعتدت علي سوريَا، كذَابون بَامتيَاز *Israel has already assaulted/attacked Syria, **excellent** liars.*

Table 4.19: Examples of negative tweets using positive words.

Positive vs. Negative

	10 Fold CV			**Ind. Test-set**			
	F	Acc.	SD	F_{pos}	F_{neg}	$F_{(pos,neg)}$	Acc.
Majority baseline (B-mjr)	0.501	64.15	0.11	0.000	0.799	0.531	66.52
Stem n-grams	0.80	80.21	1.87	0.675	0.813	0.767	76.23
Stem n-grams + Morph	**0.834**	**83.76***	1.91	0.642	**0.841**	0.774	**77.97***
Stem n-grams + Semantic	0.798	80.07	1.93	0.686	0.817	0.773	76.91
Stem n-grams + Affec-cues	0.797	79.99	1.94	0.679	0.818	0.771	76.78
Stem n-grams + Twt-topic	0.803	80.54	1.91	0.669	0.792	0.751	74.44*
Stem n-grams + Lang-style	0.795	79.75	2.19	0.584	0.794	0.724	72.47*
Stem n-grams + Twt-specific	0.795	79.80	1.99	**0.692**	0.824	**0.780**	77.60
Comb. of all feat.	0.827	83.00*	1.69	0.568	0.823	0.738	74.90*
Average	0.806	80.90	1.92	0.645	0.820	0.760	75.91

Table 4.20: Binary classification on GS1+GS2: positive vs. negative. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive Vs. Negative		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	37.483 (0.000)	0.132	0.2377
Stem n-grams + Morph	28.154 (0.000)	0.115	0.2203
Stem n-grams + Semantic	53.976 (0.000)	0.157	0.2372
Stem n-grams + Affec-cues	51.343 (0.000)	0.153	0.2418
Stem n-grams + Twt-topic	123.453 (0.000)	0.238	0.2514
Stem n-grams + Lang-style	17.031 (0.000)	0.088	0.2588
Stem n-grams + Twt-specific	33.172 (0.000)	0.123	0.2240
Comb. of all feat.	23.547 (0.000)	0.105	0.2533

Table 4.21: Comparison of performance using different feature-sets of GS1+GS2 data-set on the independent test-set.

4.5.3 Three-way classification: Positive vs. Negative vs. Negative

Results for experiments on the three-way classification task are shown in tables 4.22 and 4.23.

10 Fold Cross-Validation (CV): All classifiers significantly outperform a majority baseline. The highest performance here is achieved by the combination of all feature-sets that resulted in a significant gain of 4.9% over the stem n-grams baseline. The best individual contribution is recorded with a morphological feature-set, with a significant gain of 3.73% over stem baseline.

Independent test-set: The best performance is attained by the semantic feature-set at 64.10% accuracy, which is 0.70% improvement over the stem baseline. Table 4.23 shows that the semantic features attained the lowest classification error score at 0.358 with a small effect size at 0.17 in this set of experiments. Semantic features have also been shown to be informative for SA on Arabic newswire (i.e. MSA) [7] and SA on English tweets [187].

Overall, the combination of GS1+GS2 has again yielded an improvement on this task, attaining an average accuracy score at 61.95% as compared to the average scores attained by each of the data-sets individually on the same task, which is an improvement of 16.37% over accuracy reached with GS1 and 2.66% improvement over average accuracy of GS2. The performance gap is reduced from 36.79% on GS1 to 9.87% on GS1+GS2 for this task.

Positive vs. Negative vs. Neutral

	10 Fold CV			**Ind. Test-set**					
	F	Acc.	SD	F $_{pos}$	F $_{neg}$	F $_{neut.}$	F $_{(pos,neg)}$	F $_{(pos, neg,neut)}$	Acc.
Majority baseline (B-mjr)	0.358	52.18	0.06	0.000	0.528	0.000	0.264	0.239	41.04
Stem n-grams	0.698	70.24	1.49	0.584	0.617	0.676	0.601	0.633	63.40
Stem n-grams + Morph	0.737	73.96*	1.44	0.545	**0.640**	0.581	0.593	0.598	59.95*
Stem n-grams + Semantic	0.697	70.17	1.46	**0.591**	0.633	**0.677**	**0.612**	**0.641**	**64.10**
Stem n-grams + Affec-cues	0.701	70.52	1.45	0.588	0.625	**0.677**	0.607	0.637	63.79*
Stem n-grams + Twt-topic	0.705	70.92*	1.53	0.574	0.624	0.661	0.599	0.628	62.58*
Stem n-grams + Lang-style	0.719	72.33*	1.48	0.517	0.587	0.665	0.552	0.602	60.83*
Stem n-grams + Twt-specific	0.710	71.28*	1.53	0.572	0.616	0.643	0.594	0.617	61.67*
Comb. of all feat.	**0.749**	**75.14***	1.38	0.511	0.630	0.593	0.571	0.591	59.27*
Average	0.715	71.82	1.47	0.560	0.622	0.647	0.591	0.618	61.95

Table 4.22: Three-way classification on GS1+GS2: positive vs. negative vs. neutral. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

	Positive Vs. Negative Vs. Neutral		
	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Stem n-grams	123.926 (0.000)	0.187	0.3660
Stem n-grams + Morph	2.106 (0.349)	0.029	0.3993
Stem n-grams + Semantic	108.735 (0.000)	0.175	0.3589
Stem n-grams + Affec-cues	97.749 (0.000)	0.166	0.3615
Stem n-grams + Twt-topic	302.939 (0.000)	0.292	0.3773
Stem n-grams + Lang-style	94.784 (0.000)	0.163	0.3858
Stem n-grams + Twt-specific	104.379 (0.000)	0.172	0.3756
Comb. of all feat.	5.422 (0.066)	0.041	0.4115

Table 4.23: Comparison of performance using different feature-sets of GS1+GS2 data-set on the independent test-set.

4.6 Conclusions

The analysis of SL results is carried out throughout this chapter with respect to: the performance of different feature-sets, the overall/per-class performance, the impact of data size in closing the performance gap between CV and independent test-set settings for model evaluation and the performance of MSA vs. MSA+DA classifiers.

Observations on our GS data-sets: The overall performance indicates that the trained models can outperform a majority baseline across all feature- and data-sets.

Feature-sets: With respect to feature-sets, the stem n-grams have shown to be key features attaining up to 75.11% on binary classification (average of polarity classification at 73.99% and sentiment classification at 76.23%) and 63.40% accuracy scores on three-way classification. This is in line with previous SA results on Arabic tweets at an average accuracy of 68.35% [6]. The addition of different blocks of feature-sets to the stem n-grams baseline has generally shown a positive effect on the classifiers' performance. For instance, the morphological feature-set has resulted in improving the overall performance especially with the binary classification of positive vs. negative. That is, despite being extracted using a publicly available morphological analyser originally designed for MSA (page 62), morphological features have been shown useful for SA despite being noisy. A possible explanation is that variations of some of the morphological features (e.g., existence of a gender, person feature) may correlate more frequently with positive or negative sentiment [8]. For instance, we found that masculine form occurs more with negative tweets. The non-word based features, like semantic features, have also resulted in notable performance gains, especially for the three-way classification. This confirms that semantic features are not only useful for SA on MSA [7], but also for Arabic social media. In addition, a language-independent and non-word-based feature like Twitter-specific have been shown discriminative, e.g. re-tweet occurs more with positive tweets while hashtags tend to appear in negative tweets (section 4.4.3). Other non-word based features, like affective-cues and tweet-topic, have shown a positive, but less significant impact on the overall performance. This possibly suggests that further expanding the

affective-cues lexica and employing a more sophisticated means for topic modelling (e.g. [111]) might further enhance the effectiveness of these features. Finally, the language-style feature-set seemed to have a positive impact with the CV setting and hurt the performance with the independent test-set setting. This possibly indicates that expressive means/patterns which are employed to reflect sentiments can change over time. The result is an increased mis-match between training and testing data-sets in terms of captured patterns resulting in a negative effect for this feature-set on the independent test-set. Generally, this feature-set seems to indicate that negative instances tend to be longer and use more punctuation than positive tweets, which is also observed by Socher et al. [160] in English.

Per-class performance: The per-class results indicate that some classes are more difficult to detect than others. In particular, predicting positive tweets is found to be a more problematic task as compared to negative and neutral instances. A possible explanation that we observed is the tendency to use positive words in a negative context, either sarcastically or as a means of stressing/amplifying the intended feeling (section 4.5.2). However, the low performance on the positive class in the experiments presented in this chapter can be partially an effect of class distribution, as positive is the minority class in all of the GS data-sets and in our test-set.

MSA vs. MSA+DA: Using the B-MSA baseline has shown that the addition of dialectal tweets has resulted in a significant gain across all classification tasks. This suggests a usefulness for their presence despite the noise they introduce (e.g. for morphological features extracted using a morphological analyser developed for MSA only), which is also observed by Zbib et al. [185].

Evaluation procedure: It can be inferred that CV can be effectively used for SA applications targeting a specific time and/or constraint topic/event (e.g. political elections), while the independent test-set setting can be used for developing general-purpose applications, i.e. systems which run for longer periods of time. While it is common to observe a drop in performance between CV and an independent test-set [7], our results indicate that increasing the size of the training set can result in

significantly improving the performance on the independent test-set and reducing the performance gap notably.

The differences in sizes between folds in CV and our independent test-set can play a role in the variation of performance between the two settings. For instance, the size of our independent test-set in GS1+GS2 experiments (section 4.5) is 4.5 times the size of folds used in CV settings. In order to assess the impact of this variation, we experimented on a random sample of our test-set that is equal in size to folds used in GS1+GS2 experiments revealed a performance drop of 7.57% as compared to 9.08% obtained when using the entire test-set for binary classification. As for three-way classification, the observed performance drop with a test-set subset is at 16.84% as compared to 15.87% observed with the entire test-set. This suggests that only a small variant in performance can be attributed to the difference in evaluation-set sizes between CV and independent test-set settings.

Size of training data: The combination of GS1+GS2 data-sets has resulted in significant improvement as compared to the performance attained by each of the data-sets individually. More specifically, table 4.24 shows that the average performance gap has been reduced from 24.13% with GS1 (2.3k instances) to 7.6% with GS2 (6.8k instances), reaching only 4.9% with GS1+GS2 (9k instances). This indicates a utility for expanding the training set on the classifiers' ability to attain better scores [31]. In the next chapter, we examine the possibilities of further expanding training by exploiting existing clues (e.g. emoticons) to automatically obtain sentiment labels.

2-way vs. 3-way classification: Table 4.24 shows that there is a consistent superiority for the 2-way classification over the 3-way classification. That is, the hierarchical 2-level binary classification approach is found to outperform the single-level 3-way classification approach with an accuracy improvement of 18.16% on GS1, 13.67% on GS2 and 12.15% on GS1+GS2 data-sets. As such, in chapter 7 we present our attempt for utilising the 2-way approach in designing a tool for automatic detection of sentiment in Arabic tweet.

Data-set (size)	**Classification task**	**Average acc.**	**Performance gap** (CV - ind. test-set)	**Average gap** (2-way & 3-way)
GS1 (2.3k)	2-way	63.74	24.13	30.46
	3-way	45.58	36.79	
GS2 (6.8k)	2-way	72.96	07.63	11.66
	3-way	59.29	15.69	
GS1+GS2 (9k)	2-way	74.10	04.93	07.4
	3-way	61.95	09.87	

Table 4.24: Summary of GS results on 2-way vs. 3-way classification tasks.

Comparison with previous SA work on Arabic tweets: Our experiments show that SA classifiers trained on the M&D data-set exploiting our feature-sets has outperformed results reported in previous studies on the same data-set using 10 fold-CV setting and SVM classifiers. Specifically, our results are 2.65% accuracy better for subjectivity classification and 9.42% accuracy better for sentiment classification than those reported by Mourad and Darwish [120]. In addition, Salameh et al. [153] reported their best performance on the M&D data-set at an accuracy score that is 6.61% lower than our results on positive vs. negative task. This suggests a superiority for our features, especially the rich set of morphological features that we utilise (section 4.2).

While not directly comparable, our results on the independent test-set have also outperformed previous work employing a hold-out test-set setting (but using a different test-set). For binary classification (positive vs. negative), in [8, 6] the authors report accuracy scores of up to 69.84% while our classifiers achieved up to 75.98% on the same task.

Comparison with inter-annotator agreement: While Kappa is not directly transferable to accuracy, it is still interesting to use consensus annotation as a reference. That is, a good reference point for performance is the agreement between annotators, as it gives an indication about how difficult the task is and how well we can expect the systems to perform [145, 121]. In this work, we calculate a standard statistic Kappa as 69.0% for positive vs. negative vs. neutral (see section 3.1.4 on

page 54). The best score with three-way classification is at 64.10% accuracy, which is promising when compared to the accuracy that is expected to be reached if the sentiment labels were assigned by human annotators.

Comparison with SemEval'15: Following SemEval's setting and in addition to calculating a weighted F-score for all of the three classes $F_{(positive,negative,neutral)}$, we also calculate the averaged F-score for $F_{(positive,negative)}$, discarding F-neutral (see table 4.22). The top preforming system in 2015's competition has achieved an $F_{(positive,negative)}$ at 0.648 (with 10k English tweets) [145], while our system (with 9k Arabic tweets) has attained a comparable top $F_{(positive,negative)}$ score at 0.612. This indicates promising progress on SA for Arabic tweets given the additional challenges associated with SA on Arabic as compared to English (discussed in section 2.3.3 on page 24).

Learning curve of the gold-standard data: Figure 4.6 shows that the sentiment classifier trained on the GS data is able to consistently benefit from adding more training data. Consequently, increasing the size of training data is potentially useful for improving the performance of an SA classifier. However, continuously obtaining such data using manual annotation is costly and time consuming. In the next chapter, we explore the possibilities of employing an automatic means for obtaining large amounts of training data with no human intervention.

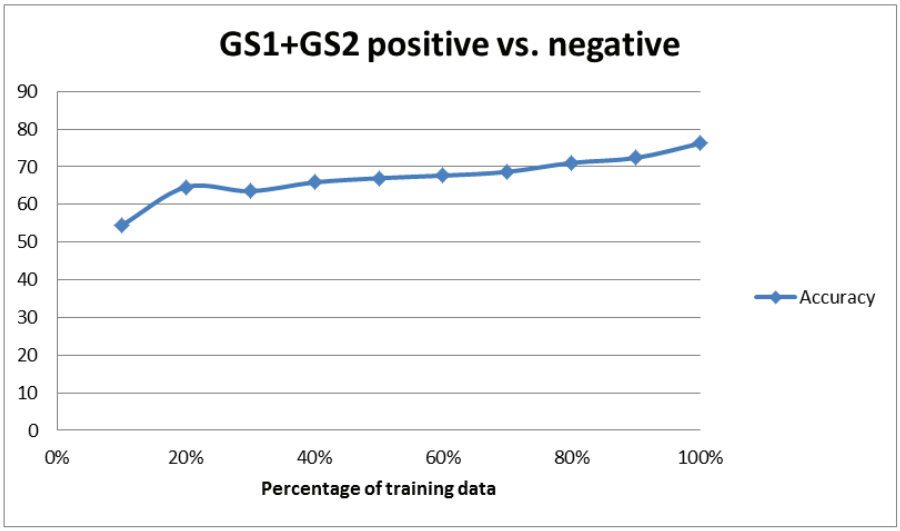


Figure 4.6: Learning curve for the gold-standard data-set GS1+GS2.

4.7 Summary

In this chapter, we empirically investigate the utility of an existing successful fully SL approach for SA on Arabic tweets. SL has been previously applied in the state-of-the-art SA systems on English tweets [123, 146, 145]. We experiment with learning classifiers on one existing and two newly collected and manually annotated corpora. The data-sets used are larger than data used in previous SA studies in Arabic tweets. Specifically, using a 9k corpus of Arabic tweets, our SA systems are able to attain a comparable performance to that attained by the state-of-the-art SA systems for tweets in English, as a well-resourced language. Our corpus has been already made publicly available to the research community.[3]

We use publicly available tools to extract features (e.g. morphological analysers and subjectivity lexica). Despite being noisy since they were extracted using resources developed for MSA only, the extracted features have been shown to remain useful/informative for the performance of SA classifiers. The best performance is recorded at 73.99% accuracy for subjectivity classification (polar vs. neutral) and at 77.97% accuracy for sentiment classification (positive vs. negative).

Our investigations also revealed that performance gap typically encountered with streaming data (e.g. topic shift issues) [82] can be reduced when using a larger

[3]Available at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`.

training data. This suggests that more data is generally beneficial for improving the generalisation ability of the classification models, i.e. reduce/eliminate overfitting [31]. However, continuously obtaining training data manually is costly. A possible solution that has been exploited in literature is by using cheap, but noisy, training data obtained using *distant supervision* approaches [81, 135].

What next? Our next step is to extend our current data-sets and empirically assess the usefulness of distant-supervision learning approaches for Arabic in order to continuously adapt to the dynamic nature of the Twitter stream.

Chapter 5

Distant Supervision Approaches

5.1 Introduction

This chapter investigates training sentiment classifiers using large automatically labelled data-sets obtained by using Distant Supervision (DS) approaches. We study their impact on the classification accuracy compared to the traditional SL approaches that use manually labelled data (chapter 4). Two DS methods for generating automatic sentiment labels are explored in this chapter. One method uses conventional markers in Twitter, i.e. emoticons and hashtags, and the other one uses existing polarity lexica, e.g. MPQA and ArabSenti. Parts of this chapter are published in [140, 139].

5.1.1 Why Distant Supervision?

Chapter 4 shows that utilising a high quality manually-annotated (gold-standard) data is beneficial for learning SA models. In addition, results indicate that adding more training data is useful to reduce/eliminate the impact of topic/temporal dependency (i.e. a performance gap resulting from model overfitting) usually encountered with streaming data (e.g. the Twitter stream) [136, 137, 37, 31, 82]. However, continuously obtaining gold-standard data to cope with the rapidly-changing nature of Twitter is unpractical and costly. So far, it is an open research question whether there can be a saturation of data [31, 82]. Since language in social media is dynamically changing and developing [61, 82], we think that new, up-to-date training data

will always be necessary. A common approach that has been successfully exploited in the literature to remedy this issue is Distant Supervision (DS) [81, 135]. DS approaches promise to remedy this overfitting by learning from very large, but noisy data.

The use of DS approaches in previous work for SA is mainly motivated by the fact that they are cheap and effective. That is, DS has been successfully used for SA in social media (e.g. Twitter) wherein raw data is freely available in large amounts but their labels are expensive to obtain [156, 147]. Therefore, advantages of the DS approaches are: first, providing alternatives to the laborious and expensive methods to obtain manual labels; Second, building larger training sets in a timely manner and hence improving coverage of lexical variations that SA classifiers can learn along with their association to a sentiment label [54]. In DS, sentiment labels are obtained using existing features, such as emoticons and sentiment-bearing hashtags, to serve as noisy labels [147].

5.1.2 What Are the Alternatives?

Existing alternative solutions to cheaply obtain training data that have been proposed in the literature include crowdsourcing and active learning.

Crowdsourcing is "the delegation of a particular task to a large group of untrained individuals rather than a select trained few" [182]. This approach has been exploited to collect training data for NLP tasks in Arabic, e.g. dialect identification [184]. However, a major challenge with crowdsourcing is annotator reliability, especially with the anonymity of individuals involved [95, 135]. Several techniques have been proposed to tackle this issue and increase quality of obtained annotations [130]. For example, 'catch trials' is one possible solution that allows identifying inattentive individuals [130]. This includes those who successively fail to provide answers matching gold-standard ones for a selected sample known only to the task creator [130]. Other issues with crowdsourcing are annotation cost [182] and ethical issues [79].

The other alternative is Active Learning (AL), where "a machine learner may

pose queries, usually in the form of unlabelled data instances to be labelled by an oracle (e.g. a human annotator)" [156]. For text classification problems, a number of issues regarding AL were identified. For instance, an active machine learner is prone to query outliers, i.e. when the least certain data instance lies on the classification hyperplane/decision-surface (page 70) [157]. Thus, knowing the label of the selected outliers is unlikely to improve accuracy of the model. Instead, it is likely to increase computational costs and training time [157]. In addition, Baldridge and Osborne [29] argue that if the model to be trained is changed, randomly labelled data can often be better than data selected in active learning with a different model. NLP researchers, however, are interested in annotating data once and use it to develop different models. Thus, Baldridge and Osborne [29] suggest adopting other "cost-saving" strategies to obtain annotated data. In this context, DS approaches have been widely adopted, showing a considerable success [136, 137, 81, 180, 135] (see section 5.2).

5.2 Related Work

This section outlines the literature related to the experimental work performed in this chapter. It reviews previous work thematically, based on the methods used, i.e. features exploited to automatically obtain sentiment labels.

5.2.1 Conventional-Markers-based DS Approach

In this section, we concentrate on previous work that has used emoticons and/or hashtags to automatically annotate a Twitter data-set, which is then used to train a sentiment classifier.

An early attempt to study sentiment classification in Twitter exploiting the emoticons used in Twitter messages was made by Go et al. [81]. Using emoticons to build a training data-set of English tweets, the authors trained several machine learning classifiers (SVM, MaxEnt, and NB) to perform a binary (positive vs. negative) classification. The final training-set comprises 1.6M English tweets with an equal number of positive and negative instances, i.e. balanced data-set. For feature-

sets, they used POS and word-based 1g and 2g. The best reported result on a manually annotated test-set of 359 instances is at 83% accuracy.

Subsequent work by Bifet and Frank [37] carried out SA on English tweets using an automatically labelled training data-set using emoticons. Unlike Go et al. [81], Bifet and Frank [37] experimented with balanced vs. unbalanced classes. The reason is that the authors argue that a representative sample of training data from the Twitter stream is less likely to be balanced. They used word-based 1g as features. Training an NB sentiment classifier and testing it on the same test-set used by Go et al. [81] yielded an accuracy score of 82.45%. Their experiments on a highly unbalanced data-set (predominantly with positive tweets) yielded accuracy scores of up to 73.81%.

Similarly, Pak and Paroubek [127] used emoticons to collect an English Twitter corpus and build a sentiment classifier. Unlike the binary sentiment classification used by Go et al. [81] and Bifet and Frank [37], they expanded the scope of investigations to perform three-way classification positive vs. negative vs. neutral. For automatically building a neutral training corpus, they collected a set of neutral instances from Twitter accounts of popular newspapers, e.g. New York Times. For feature-sets, they used word-based 1g, 2g and 3g. Evaluating their models on a manually annotated test-set of 216 tweets, their experiments yielded an F-score at 0.60 with an NB classifier. We follow their idea of collecting neutral Twitter messages from popular news accounts.

As for hashtags, Kouloumpis et al. [109] used sentiment-bearing hashtags (e.g. #fail, #job) to automatically annotate a set of English tweets to experiment on three-way classification positive vs. negative vs. neutral. The authors evaluated the trained classifiers on a manually annotated test-set. They report the best results when combining a set of syntactic, semantic and stylistic features at 74% accuracy and 0.68 F-score.

DS has also been explored for emotion analysis, i.e. classifying emotion types like happy, sad, anger, etc. In this context, Purver and Battersby [135] empirically investigated the performance of supervised classifiers trained with an automatically

labelled training data-set (using emoticons and hashtags) to perform multi-class emotion analysis on English tweets. The authors found DS approach to be more reliable for detecting: happiness and sadness (which corresponds to the positive/negative sentiment classification [23]) with the best reported F-score for evaluating the trained SVMs on a manually annotated test-set being 77.5% for happiness and 54.5% for detecting sadness with the emoticons data-set, and 62.6% for happiness and 60.4% for sadness with the hashtags data-set. It is interesting to see that even with a more fine-grained emotion analysis task, happiness (positive) and sadness (negative) seem to be amongst the most distinguishable emotions.

A subsequent study by Suttles and Ide [163] has also used an automatically labelled data-set of English tweets to perform emotion analysis. For sentiment annotation, they used emoticons, hashtags and emoji. They trained MaxEnt and NB classifiers using word-based 1g as features. Evaluating the models on a manually annotated test-set, the authors reported an accuracy score of up to 90.6% for binary classification task discriminating joy/sadness instances.

Similarly, for determining emotion type for less-resourced languages, such as Chinese, Yuan and Purver [180] performed experiments to detect emotions from a Chinese micro-blog service. Emoticons were used to generate emotion labels for six emotion classes. By training SVM classifiers, the authors reported that happiness is the most discriminative class with an accuracy score of up to 78.2%, followed by sadness at an accuracy score of 69.6%.

As for Arabic, AlMutawa [22] describes a number of experiments the author conducted to carry out emotion analysis on a balanced data-set of Arabic tweets that were automatically labelled for six classes of emotion using emoticons and hashtags. For features, the author used word-stem 1g, 2g and 3g. By training SVM classifiers on an emoticon labelled data-set, the author evaluated the trained models on a manually annotated test-set reporting accuracy scores at 57.69% for detecting happiness and 45% for sadness (average accuracy is 51.35%). For the hashtag-based data-set, the attained accuracy scores were up to 63.42% for happiness and 70% for sadness (average accuracy is 66.71%). Our work is different in investigating

the use of emoticons and hashtags for determining sentiment polarity (positive or negative) rather than emotion type (happy, sad, anger, etc.). In addition to word n-grams, we explore a wide set of features and evaluate the trained models against an independent test-set.

Summary: To the best of our knowledge, no previous work other than [22] has investigated the utility of exploiting conventional markers, i.e. emoticons and/or hashtags, to be used as noisy labels to mark the emotional orientation of authors in Arabic social media posts. DS is expected to be especially promising for Arabic as larger data-sets are required to enhance vocabulary coverage. Thus, this possibly helps to overcome/alleviate the impact of Arabic's morphologically-rich nature, i.e. many word forms, and presence of DAs, i.e. different dialects use different expressions to deliver the same sentiment (page 12).

5.2.2 Lexicon-based DS Approach

In this section, we review previous studies that have considered a lexicon-based SA approach to either ultimately determine the sentiment orientation of a given text instance or automatically obtain sentiment labels for training instances, which were then used to train ML sentiment classifiers.

5.2.2.1 A Lexicon-based Approach for SA

Read and Carroll [137] present a study investigating the effectiveness of three sentiment dictionaries compiled using lexical-association/word-similarity methods, e.g. Pointwise Mutual Information (PMI). The dictionaries were built using large news-based corpus and tested on a corpus of English movie reviews. The sentiment of each review is determined as the sign of the sum (+/-) of the sentiment scores for each extracted sentiment-bearing word. The authors perform a binary classification (positive vs. negative) and report the best F-score at 0.687.

A succeeding study by Taboada et al. [165] presented a lexicon-based system for SA on English. The proposed system incorporates semantic orientation of individual words and contextual shifters (i.e. negators). Unlike Read and Carroll [137], the

authors here created a manually annotated sentiment dictionary that includes nearly 5k words, with each of the words being assigned with a hand-ranked sentiment orientation value (positive or negative). Neutral words were excluded from the final dictionaries. For negation handling, the authors considered two alternative methods. The first one is negation switch in which the polarity score will be shifted once a negator is detected. The second method is negation shift in which the sentiment score will be adjusted with a fixed amount. Performing binary classification (positive vs. negative), they report accuracy scores of up to 78.74% on reviews and 75.31% on blog posts.

On English tweets, Thelwall et al. [167] presented a lexicon-based system for SA that utilises, besides sentiment dictionaries, a set of rules to detect the strength of a sentiment in short informal English text instances. The system is called *SentiStrength*.[1] Unlike the system proposed by Taboada et al. [165], which calculates a single polarity score indicating the overall polarity of a given text instance, SentiStrength calculates two scores: one for positive class and another for negative class (both ranging from 1=neutral/no-sentiment to 5=strong sentiment), assuming the coexistence of positive and negative sentiment even in short text instances (e.g. tweets). In addition to semantic rules previously used by Taboada et al. [165] (e.g. handling negation), Thelwall et al. [167] enhanced their system with other components as a special adaption to the social web genre, i.e. accounting for the presence of emoticons, repeated punctuations, repeated letters and all capitalised words. To evaluate their system, they used several manually annotated data-sets representing various domains, such as BBC Forum posts, tweets and YouTube comments. The authors reported an accuracy score of 62.65% for detecting positive and negative tweets.

On Italian, as a less-resourced language, Basile and Nissim [34] described their attempts for carrying out SA on Italian tweets. Due to lack of resources available, the authors created a new Twitter data-set and a sentiment lexicon. The Twitter data-set was manually annotated by native speakers, while the sentiment lexicon was obtained by mapping Italian synsets in MultiWordNet to the sentiment score

[1] `http://sentistrength.wlv.ac.uk/`

of their corresponding English synset in SentiWordNet. To determine the overall sentiment orientation of a given tweet, their system sums the polarity scores of captured word-tokens. Performing a three-way classification (positive vs. negative vs. neutral), the authors reported the best scores at an F-score of 0.495 and accuracy of 55.4%.

As for Arabic, an investigation was conducted by Albraheem and Al-Khalifa [20] using a small set of 100 manually annotated tweets focusing on the Saudi dialect. The authors also manually built a sentiment lexicon of positive and negative words used in this dialect. The sentiment orientation of a given tweet is determined based on the sum of polarity words identified. Using this method, the authors reported an accuracy score of 73% on positive vs. negative classification. Again, this probably suggests that a better performance can be achieved with SA systems tuned for a specific dialect.

A subsequent study by Abdulla et al. [9] performed a lexicon-based binary sentiment classification on a data-set of MSA and Jordanian tweets. The authors used a manually annotated lexicon to extract sentiment-bearing words from given tweets and assign tweets with an aggregated sentiment score, which is then used to assign tweets with sentiment labels, i.e. based on the sign of the tweet's score. Comparing the auto-generated labels to manually assigned ones, the authors reported an accuracy of 59.6%.

Another study by El-Beltagy and Ali [62] carried out SA on a data-set of 500 Egyptian tweets. The lexicon was created by manually annotating a small set of words, which were then used as seeds to automatically expand the sentiment lexicon. The resultant lexicon was ultimately manually filtered to exclude irrelevant entries. Each entry is associated with a sign +/- indicating its sentiment orientation and a strength value. The sentiment labels were assigned to tweets based on the sign of the added up score of extracted positive/negative words. Similar to Taboada et al. [165], the authors experimented with a uniform weighting scheme, i.e. positive word= +1 and negative word= -1, for which they reported an F-score at 0.496 for discriminating positive and negative instances. In addition, The authors experimented using a

weighting scheme exploiting the strength value of each detected sentiment-bearing word from the lexicon reporting an F-score at 0.702.

In a recent study by Wang et al. [171], the authors developed a system for SA on Arabic tweets exploiting a sentiment lexicon that was translated from English and manually filtered. The scope of the study focused on Egyptian and Saudi dialects. Therefore, the authors expand the lexicon by manually adding the Egyptian and Saudi equivalent of each entry in the lexicon. In addition, they restrict their Twitter data-set to cover only three topics: Egyptian government, telecommunication and employment. Testing the proposed system on manually annotated tweets, the average recorded F-score is at 80.14%.

5.2.2.2 A Combined Approach for SA: Lexicon-based + Machine-Learning

Zhang et al. [186] employed a hybrid approach on a data-set of English tweets. First, they apply a lexicon-based approach, which uses a publicly available opinion lexicon to determine the sentiment orientation for each tweet. Then, they use the annotated examples to train a sentiment classifier to perform a three-way classification (positive vs. negative vs. neutral) using word-based 1g as features. The trained SVM was tested against a random sample of manually annotated tweets reporting an accuracy score up to 85.4%, which is encouraging to apply this approach on Arabic tweets.

With regard to Arabic, a recent study by El-Makky et al. [64] used a hybrid approach similar to that employed by Zhang et al. [186]. The authors carried out SA on a data-set of Egyptian tweets, similar to El-Beltagy and Ali [62]. For the sentiment lexicon, they used publicly available lists (i.e. MPQA and ArabSenti) and an in-house lexicon. The resultant lexicon was then used to calculate the sentiment orientation score for each given tweet. They trained a sentiment classifier exploiting various feature-sets including: POS, Twitter-specific, language-style (e.g. presence of elongation). With a 10-fold CV setting, the authors reported F-scores at 0.72 for subjectivity classification and 0.79 for sentiment classification.

Conclusion: There have been limited attempts to apply DS approaches to Arabic. Unlike previous work, we investigate the utility of DS approaches on multi-dialect

Arabic Twitter corpora. We also systematically compare DS approaches against the results obtained with SL approaches (chapter 4) by benchmarking them against our independent test-set to gain insights about the aspect of data quality vs. quantity for SA.

5.3 DS Experiments: Part One

The aim of this set of experiments is to investigate the usefulness of a number of DS methods, exploiting large (but noisy) data, for SA annotation in Arabic tweets. We evaluate the approaches against our independent test-set (page 54), following Go et al. [81]. This will allow a direct comparison to the best recorded scores attained by our fully SL approach that uses a smaller (but gold-standard) data (see chapter 4).

Part one covers investigations using three DS-based data-sets: emoticon-based (Emo1), lexicon-presence-based (lex-Pres1) and lexicon-aggregation-based (lex- Aggreg1) data-sets (page 47). The data-sets are used to perform two binary classifications (i.e. for subjectivity and sentiment) and a three-way classification (positive vs. negative vs. neutral). Results in part one of the experiments indicate that the three DS data-sets generally perform better than a fully SL approach for subjectivity classification (polar vs. neutral), but perform less effectively on sentiment classification (positive vs. negative). To investigate possible ways for boosting the performance of sentiment classifiers using DS methods, we conduct a second round of experiments (section 5.4).

5.3.1 Experiments on the Emoticon-based (Emo1) Data-set

In this section, we evaluate the potentials of exploiting training data that is automatically labelled using (noisy) emoticons, see page 55. As explained on page 48, the neutral tweets in all DS experiments were automatically collected using official news accounts on Twitter (e.g. BBC-Arabic). The results of this set of experiments are summarised in Table 5.1.

Polar vs. Neutral: The average accuracy score is at 94.78%. This is an improvement of 22.49% over the average accuracy score attained at at 72.29% by the best fully-supervised system, i.e. trained on GS1+GS2 data-set. It is worth mentioning that the neutral class in the independent test-set are randomly collected and manually annotated and they may/may not include news tweets. These results indicate that the classifier is able to effectively recognise and distinguish the language

Emo1 Data-set

	Polar vs. Neut.		**Pos. vs. Neg.**		**Pos. vs. Neg. vs. Neut.**	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.471	61.70	0.531	66.51	0.239	41.04
Stem n-grams	0.949	94.89	0.50	50.29	**0.704**	**69.67**
Stem n-grams + Morph	**0.950**	**95.19**	0.510	51.25	0.690	68.43
Stem n-grams + Semantic	0.940	94.28	**0.540**	**53.93**	0.675	67.18
Stem n-grams + Affec-cues	0.907	93.27	0.527	53.26	0.584	61.33
Stem n-grams + Lang-style	0.930	92.80	0.531	53.58	0.597	61.71
Stem n-grams + Twt-specific	0.937	94.59	0.510	50.97	0.697	69.37
Comb. of all feat.	0.942	94.74	0.530	53.67	0.619	63.41
Average	0.948	94.78	0.541	53.97	0.652	65.87

Table 5.1: Binary and three-way classification on Emo1 data-set: polar vs. neutral, positive vs. negative and positive vs. negative vs. neutral.

used to express neutral/objective utterances from those used to convey personal opinion/attitude.

Positive vs. Negative: The average accuracy is at 53.97%, which is 21.94% lower than the accuracy score attained by the GS1+GS2 data-set at 75.91% on the same task. This indicates a high level of noise introduced with using emoticons as noisy labels. Later in this chapter, we will investigate possible reasons for this performance drop in a detailed error analysis (section 5.4).

Positive vs. Negative vs. Neutral: When performing three-way SA, the classifiers achieve an average accuracy of 65.87% accuracy, which is 3.92% improvement over the accuracy score attained by GS1+GS2 at 61.95% on the same task. The confusion matrix reveals that detecting the positive and negative is the most problematic here, and that detecting neutral boosts performance the most. This is clearly reflected in the recorded per-class F-scores at 0.535 for $F_{positive}$, 0.466 for $F_{negative}$ and 0.942 $F_{neutral}$.

5.3.2 Experiments on the Lexicon-presence-based (Lex-Pres1) Data-set

In this section, we experiment with a lexicon-presence-based approach to DS. That is, instead of using emoticons, we now utilise a combined sentiment lexicon to automatically assign noisy sentiment labels (details on the creation and annotation of the lexicon-based data-sets, see page 51). In this setting, the sentiment labels are automatically assigned to tweets based on the presence of positive/negative sentiment-bearing words, with tweets including mixed emotions being excluded (table 3.8 on page 55). Results of this set of experiments are summarised in table 5.2.

Polar vs. Neutral: The attained average accuracy score at 95.26% surpasses that recorded with GS1+GS2 data-set at 72.29% on this task by an improvement of 22.97%. The accuracy here is almost identical to that achieved by the emoticon-based approach on this task at 94.78%. This experiment utilises the same neutral set used in Emo1 experiments.

Positive vs. negative: Again, we note that it is difficult to discriminate positive vs. negative instances using this lexicon-presence-based DS approach. That is, the average accuracy score with Lex-Pres1 data-set is at 55.51%, which is 20.4% lower than that achieved by GS1+GS2 on this task at 75.91%. Compared to Emo1 results, it is interesting to note that the lexicon-presence average accuracy score is 1.54% better than the emoticon-based approach on this task at 53.97%, which allows us to infer that lexicon-presence labelling might introduce less noise for SA.

Positive vs. negative vs. neutral: The average accuracy score here is at 70.82%, which is 8.87% better than that achieved with GS1+GS2 on this task at 61.95%. Similar to the emoticon-based approach, the per-class performance indicates a superiority for detecting neutral class, as $F_{neutral}$ is at 0.926 as compared to 0.599 for $F_{positive}$ and 0.407 for $F_{negative}$. Zhang el al. [186] also found that 3-way classification (at 0.854 F-score) is better than the binary classification of positive

Lex-Pres1 Data-set

	Polar vs. Neut.		**Pos. vs. Neg.**		**Pos. vs. Neg. vs. Neut.**	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.471	61.70	0.531	66.51	0.239	41.04
Stem n-grams	**0.953**	**95.36**	0.530	**56.10**	**0.71**	**71.57**
Stem n-grams + Morph	0.951	95.17	0.520	55.51	0.680	70.07
Stem n-grams + Affec-cues	0.939	93.74	0.524	52.26	0.613	65.85
Stem n-grams + Lang-style	0.945	94.53	**0.544**	54.86	0.698	70.10
Stem n-grams + Twt-specific	0.941	94.52	0.539	55.46	0.691	69.58
Comb. of all feat.	0.946	94.83	0.534	54.73	0.686	69.83
Average	0.951	95.26	0.52	55.51	0.695	70.82

Table 5.2: Binary and three-way classification on Lex-Pres1 data-set: polar vs. neutral.

vs. negative (at 0.749 F-score) and attributed the difference to the superiority in detecting the instances of neutral class.

It is interesting to note that stem-ngrams features seem to be the most informative for the classifiers in all classification tasks in LexPres1 and Emo1 experiments. This possibly reflects that the auto-labelled instances hold inconsistent patterns that the extracted features represent. In section 5.4.2, we will see that increasing the size of the training data will help the classifiers to overcome such noise, making them able to capture some consistency and benefit from features other than stem n-grams.

5.3.3 Experiments on the Lexicon-aggregation-based (Lex-Aggreg1) Data-set

This section presents the experiments we have conducted on the lexicon-aggregation-based (Lex-Aggreg1) data-set (see page 55). The sentiment labels are automatically assigned to tweets based on the sign/orientation (i.e. + or -) of the summed up score of sentiment-bearing words captured in each tweet (see page 51). Results are summarised in table 5.3.

Polar vs. Neutral: The neutral instances here are the same used in Emo1 and Lex-Pres1 experiments. The average accuracy scores is at 86.47%. Similar to Emo1

Lex-Aggreg1 Data-set

	Polar vs. Neut.		**Pos. vs. Neg.**		**Pos. vs. Neg. vs. Neut.**	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.471	61.70	**0.531**	**66.51**	0.239	41.04
Stem n-grams	**0.910**	**91.11**	0.52	52.98	**0.630**	**64.96**
Stem n-grams + Morph	0.810	81.83	0.501	50.83	0.610	63.36
Stem n-grams + Affec-cues	0.849	85.06	0.510	51.20	0.625	62.23
Stem n-grams + Lang-style	0.891	88.69	0.521	52.13	0.627	62.31
Stem n-grams + Twt-specific	0.897	89.30	0.503	50.01	0.627	63.02
Comb. of all feat.	0.850	85.43	0.482	48.83	0.629	63.81
Average	0.860	86.47	0.50	50.81	0.620	64.16

Table 5.3: Binary and three-way classification on Lex-Aggreg1 data-set: polar vs. neutral, positive vs. negative and positive vs. negative vs. neutral.

and Lex-Pres1 on this task, the results here also indicate a better performance compared to the fully-supervised approach with an improvement of up to 9.54% in average accuracy over the results achieved by GS1+GS2 data-set at 72.29%. However, the performance on this data-set is generally 8% lower than average accuracy achieved by the Lex-Pres1 and Emo1.

Positive vs. Negative: The average accuracy score attained by Lex-Aggreg1 is at 50.81%, which is lower than those achieved by GS1+GS2 at 75.91%, Emo1 at 53.97% and Lex-Pres1 at 55.51% on this task. This suggests lexicon-aggregation DS approach as a less effective compared to approaches investigated by far. A possible explanation is the notable presence of mixed instances with this approach (further discussion in section 5.4.3).

Positive vs. Negative vs. Neutral: The average recorded accuracy score is at 64.16%, which is 2.21% better than that achieved by GS1+GS2 on this task at 61.95%. Both of these accuracy scores are lower than those achieved on this task by Emo1 at 65.87% and Lex-Pres1 at 70.82%.

5.3.4 Summary of Part One Results

- For polar vs. neutral classification, the results show a significant improvement over the classifier trained using a fully-supervised approach on a gold-standard data-set (i.e. GS1+GS2). We achieve the best performance on this task with the Lex-Pres1 data-set at an average accuracy of 95.26% on the independent test-set, which is a 22.97% improvement over GS1+GS2 results at 72.29%. Both Emo1 and Lex-Aggreg1 were able to outperform GS1+GS2 on this task, achieving accuracy scores at 94.78% with Emo1 and 86.74% with Lex-Aggreg1. This suggests a performance benefit that this task (polar vs. neutral) has gained with the DS approaches, which can be attributed to two main factors. First, the increase in the size of the training set. For example, Emo1 data-set in this task is around 13.2 times larger than GS1+GS2 data-set (see table 3.8 on page 55), and thus the emoticon-based model better generalises to unseen events. Second, neutral instances in the DS data-sets were sampled from news accounts, which are mainly written in MSA, whereas we hypothesise that tweets including emoticons (which we use for acquiring polar instances) are mainly written in DA (see diagram 5.1 on page 135). This has possibly caused the classifiers to learn to discriminate DA vs. MSA instead of polar vs. neutral. To investigate this hypothesis, we study the correlation between the automatically detected language class, i.e. MSA/DA, for a given tweet using AIDA (see page 68) and the accuracy of predicting this tweet. The results show a significant correlation (Pearson = 0.202, p-value=0.000) and thus confirm our hypothesis that our classifiers learn to detect MSA vs. DA.

- For positive vs. negative classification, surprisingly, none of the three DS methods are able to outperform the scores attained by GS1+GS2 on this task at 75.91% on the test-set. Among the DS methods, Lex-Pres1 reaches the best score on this task at 55.51%. Despite the fact that the Lex-Pres1 is about 6.3 times larger than GS1+GS2 (see table 3.8), this is still not enough to compete with the fully-supervised method using GS1+GS2 data-set on this task. This might suggest a certain degree of noise introduced with the DS methods

when assigning the positive/negative labels. In section 5.4, we will be further investigating reasons for this relatively poor performance and possibilities for further boosting the performance on positive vs. negative classification task.

- For positive vs. negative vs. neutral classification, all of the three DS methods outperform the score attained by GS1+GS2 on this task at 61.95%. The DS methods reach average accuracy improvements of 3.92% with Emo1, 8.87% with LexPres1 and 2.21% with LexAggreg1. The per-class scores indicate that positive and negative classes are the most problematic. For instance, the per-class F-scores recorded with Lex-Pres1 on this task are 0.599 for $F_{positive}$, 0.407 for $F_{negative}$ and 0.926 for $F_{neutral}$.

- Among the three DS approaches, it seems that Emo1 and Lex-Pres1 are performing equally well with both being able to outperform Lex-Aggreg1 in all classification tasks. A possible explanation is that mixed instances (i.e. tweets with positive and negative indicators) are excluded from both Emo1 and lex-Pres1, while in lex-Aggreg1, mixed instances are included in the data-set because positive and negative lexicon are both contribute the aggregated sentiment score (section 5.4.3).

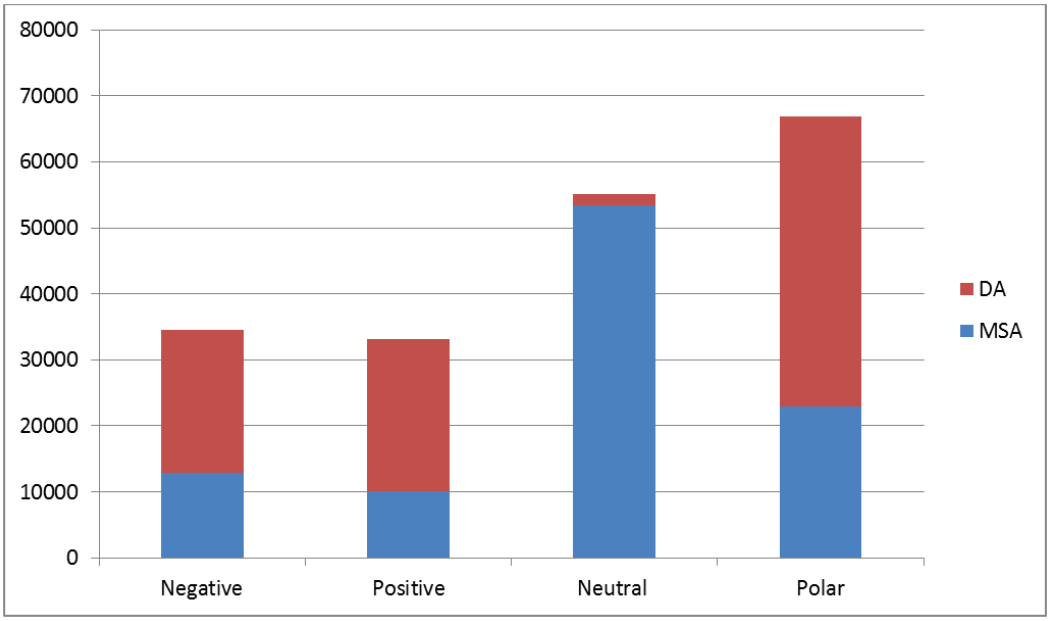


Figure 5.1: Class distribution in Emo1 data-set.

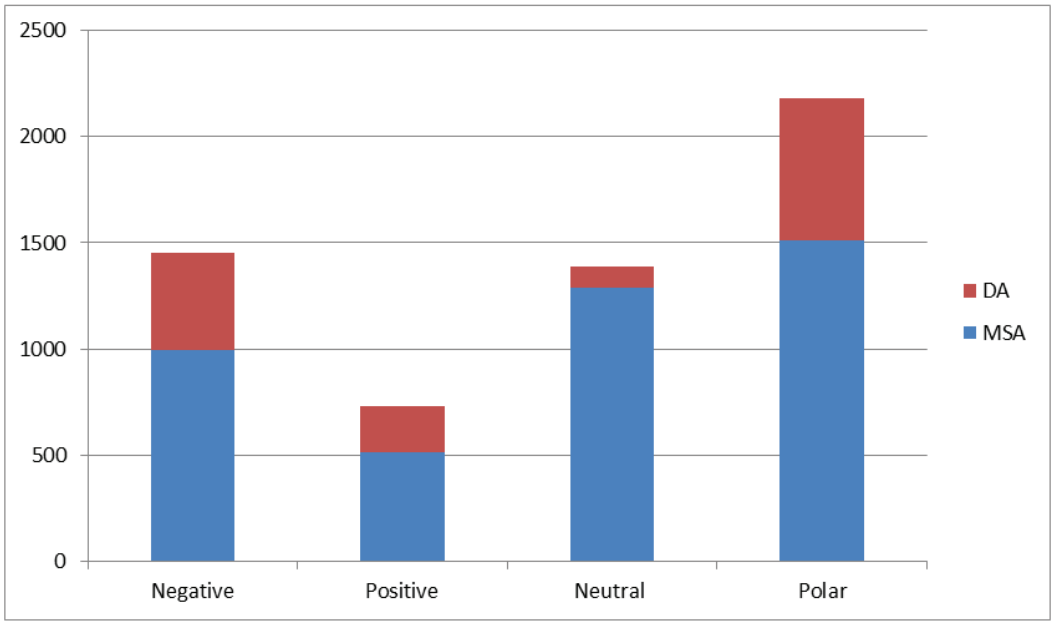


Figure 5.2: Distribution of MSA/DA instances within each class of the independent test-set.

5.4 DS Experiments: Part Two

This section focuses on the binary classification task of positive vs. negative. Results in section 5.3 show that binary sentiment classification (positive vs. negative) is a challenging task. Following previous work [136, 137, 165, 81], this section investigates possibilities for further improving the performance on this particular task, considering the following aspects: the size of training data and the use of another conventional marker (i.e. hashtags). Finally, we conduct an error analysis in order to investigate potential sources of errors.

5.4.1 Experiments on Emoticon-based (Emo2) and Hashtag-based (Hash) Data-sets

In this section, we evaluate the potentials of exploiting an extended training data that is automatically labelled using emoticons, namely Emo2 (see page 55). Emo2 data-set is 9.6 times larger than Emo1 data-set. In addition, following [186, 163, 135], we also utilise some sentiment-bearing hashtags to query emotional tweets and collect a new data-set, namely, Hash (see page 55). The results are summarised in Table 5.4.

Emoticon-based (Emo2) data-set: The results indicate an improvement of 2.56% accuracy in the overall performance when using Emo2 as a 9.6 times larger than Emo1 at 53.97%. The stem n-grams baseline has reached an accuracy score of 52.77%, which is very close to that achieved with Emo1 at 51.25%. The morphological feature-set achieved the highest performance at 64.82%. Table 5.5 shows that morphological features attained the lowest classification error rate at 0.349 with a medium effect size at 0.44. However, the results show that the best performing score is still below a majority baseline. This suggests that more data is required to boost the performance and compete the majority baseline, especially with the positive impact noted with Emo2 (as a larger emoticon-based training data).

Hashtag-based (Hash) data-set: The best performance is recorded at an accuracy score of 69.58%, which is 4.76% better than the best score achieved with Emo2. Table 5.5 shows that the lowest classification error rate is attained with the affective-cues feature-set at 0.304 with a small effect size of 0.26. This is interesting, considering the difference in size between Hash and Emo2 data-sets: Emo2 is about 8.6 times larger in size than Hash data-set. Despite that, the hashtag-based data-set is still able to notably outperform both emoticon-based data-sets (Emo1 and Emo2) on this task. A similar observation is also reported by AlMutawa [22] on Arabic tweets. Unlike Emo2, the stem n-grams baseline significantly outperforms a majority baseline. The average accuracy score of Hash is 0.30% better than Emo2. This suggests that emoticons are more noisy, and hence, less reliable, as compared to hashtags in this context. In other words, it seems that the hashtag-based data-set is less ambiguous (further discussion in section 5.4.1.1).

Combined Emo2+Hash data-set: The results show an improvement with the combination of Emo2 and Hash data-sets with an average accuracy score at 62.22%, which is 6% better than that achieved with the two data-sets individually at 56.23% with Emo2 and 56.53% with Hash. As for the stem n-grams baseline, the score attained with Emo2+Hash is at 62.81%, which is below that achieved with Hash data-set on its own at 69.22%, but is generally better than the score attained by

Positive vs. negative

	Emo2		Hash		Emo2+Hash	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.531	66.51	0.531	66.51	0.531	66.51
Stem n-grams	0.537	52.77	**0.674**	69.22	**0.621**	62.81
Stem n-grams + Morph	**0.590**	**64.82***	0.444	47.78*	0.605	**64.50***
Stem n-grams + Semantic	0.525	51.17*	0.576	62.30*	0.608	61.29
Stem n-grams + Affec-cues	0.527	51.44*	**0.674**	**69.58**	0.614	61.75*
Stem n-grams + Lang-style	0.545	53.87*	0.591	58.27*	0.568	60.19*
Stem n-grams + Twt-specific	0.555	55.15*	0.499	51.63*	0.620	62.44
Comb. of all feat.	0.531	64.41*	0.258	36.97*	0.565	62.53
Average	0.544	56.23	0.531	56.53	0.60	62.22

Table 5.4: Binary classification positive vs. negative on the emoticon and hashtag-based data-sets. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

Emo2 on its own at 52.77%. Table 5.6 shows that the morphological features in this set of experiments attained the lowest classification error rate at 0.355 with a medium effect size at 0.358. In sum, the average score across feature-sets indicates a superiority for the combined Emo2+Hash data-set, while the stem baseline indicates a superiority for the Hash data-set when used individually.

5.4.1.1 Error Analysis for Emoticon-Based DS Data-set

We conduct an error analysis in order to further investigate the underlying cause for noise with emoticon-based DS data-sets. In particular, we investigate the use of sarcasm and the direction of facing of emoticons in right-to-left alphabets.

Use of sarcasm and irony: Using emoticons as labels is naturally noisy, since we cannot know for sure the intended meaning the author wishes to express. This is especially problematic when emoticons are used in a sarcastic way, i.e. the emoticon used is different from the intended emotion [100, 135]. For instance, positive emoticons can be used in a negative context, and vice versa (see examples #1 and #2 on page 140 from our emoticon-based data-sets). This is also noted by Itani et al. [100] on a data-set of Arabic Facebook posts. Alternatively, emoticons themselves can be used truthfully, i.e. to help the reader understand that the accompanying text is being used sarcastically (see example #3). Research in psychology shows

	Emo2			**Hash**		
	χ^2 ***(p-value)***	***Effect size***	***Class. error***	χ^2 ***(p-value)***	***Effect size***	***Class. error***
Stem n-grams	54.644 (0.000)	0.159	0.4722	118.466 (0.000)	0.233	0.3078
Stem n-grams + Morph	437.091 (0.000)	0.447	0.3495	2130.402 (0.000)	0.988	0.5222
Stem n-grams + Semantic	923.25 (0.000)	0.650	0.5249	320.896 (0.000)	0.383	0.3770
Stem n-grams + Affec-cues	1045.56 (0.000)	0.691	0.5373	156.671 (0.000)	0.267	0.3041
Stem n-grams + Lang-style	105.048 (0.000)	0.219	0.4690	57.359 (0.000)	0.162	0.4173
Stem n-grams + Twt-specific	491.799 (0.000)	0.474	0.4814	1703.149 (0.000)	0.883	0.4837
Comb. of all feat.	547.607 (0.000)	0.501	0.3628	2180.08 (0.000)	0.999	0.6303

Table 5.5: Comparison of performance using different feature-sets on Emo2 and Hash data-sets.

	Emo2+Hash		
	χ^2 ***(p-value)***	***Effect size***	***Class. error***
Stem n-grams	19.752 (0.000)	0.093	0.3719
Stem n-grams + Morph	280.042 (0.000)	0.358	0.3550
Stem n-grams + Semantic	8.690 (0.003)	0.064	0.3871
Stem n-grams + Affec-cues	4.543 (0.033)	0.047	0.3825
Stem n-grams + Lang-style	193.841 (0.000)	0.297	0.3981
Stem n-grams + Twt-specific	9.510 (0.002)	0.067	0.3756
Comb. of all feat.	435.197 (0.000)	0.446	0.3747

Table 5.6: Comparison of performance using different feature-sets on Emo2+Hash data-set.

that up to 31% of the time, emoticons are used sarcastically [176]. In order to investigate this hypothesis, we manually labelled a random sample of 303 misclassified instances for neutral, positive, negative, as well as sarcastic, mixed and unclear sentiments (see Table 5.7).[2] Interestingly, the tweets identified as sarcastic represent only 4.29%, while tweets with mixed (positive and negative) sentiments represent 5.94% of the manually annotated sub-set. In 34.32% of the instances, the manual labels have matched the automatic emoticon-based labels. Surprisingly, automatic emoticon-based labels contrast with the manual labels in 36.63% of the instances. We therefore investigate a different source of error in the next paragraph. The examined test sample also includes 4.95% instances labelled as neutral. The rest of the instances were assigned 'unclear sentiment orientation'.

1 اليوم الفَالنتين جتنَا نيله في حظنَا الهبَاب :ح

It is Valentine, poor me :)

2 يمكنك التقدم بطلب لّلانظمَام الي دَاعش عبر هذَا الرَابط :ح

You can send your application to join ISIS to this link :)

3 جميل يَا اهلي :ك

great job Ahli :(*referring to a famous football team.*

Facing of emoticons: We therefore investigate another possible noise/error source following Mourad and Darwish [120], who claim that the right-to-left alphabetic writing of Arabic might result in emoticons being mistakenly interchanged while typing. On some Arabic keyboards, we observed that typing ")" will produce the opposite "(" parentheses. The examples #4 to #8 illustrate a case of misclassified instances, where we assume that the facing of emoticons might have been interchanged or mistyped. In this context, AlMutawa [22] and Al-Osaimi and Badruddin [16] also observed cases wherein the emotion being conveyed is different from the emoticon associated with the tweet, especially with happy and sad emoticons. We notice that most of these mislabelled tweets are automatically classified as positive, i.e. accompanied with positive emoticons. We therefore anticipate that, although positive is the majority class in the Emo2 data-set, the positive tweets are highly noisy.

[2]We experimented with a 10:90 test:train split of the emoticon-based data-set in order to identify possible causes for errors within the automatically obtained sentiment labels.

Emoticon Label	Predicted label	Manual label	# instances
Positive	Negative	Mixed	8
Negative	Positive	Mixed	10
Positive	Negative	Negative	59
Negative	Positive	Negative	42
Positive	Negative	Neutral	29
Negative	Positive	Neutral	7
Positive	Negative	Positive	62
Negative	Positive	Positive	52
Positive	Negative	Sarcastic	8
Negative	Positive	Sarcastic	5
Positive	Negative	Unclear sentiment indicator	19
Negative	Positive	Unclear sentiment indicator	2

Table 5.7: Results of labelling sarcasm, mixed emotions and unclear sentiment for misclassified instances.

4 خلَاص مَافي امل :ح

No hope anymore :)

5 اكرهك :ح

I hate you :)

6 تعبت وَانَا اتخيل ابي شي يصير وَاقع :ح

I'm tired of dreaming, I want something to become true :)

7 البقَاء لّله :ح اللّهم ارحمهم

Condolences :) May Allah shower their souls with mercy

8 تسلم يَاخي :ك

God bless you my friend :(

Conclusion: Generally, the results with both emoticon-based data-sets (Emo1 and Emo2) are about 18% lower than results in previous work on emoticon-based binary sentiment classification on English tweets by Go et al. [81] and Bifet and Frank [37], which both achieved around 83% accuracy. Whereas, previous emoticon-based work on Arabic reported an average accuracy of 51.35% for detecting happy/sad tweets [22]. This indicates, together with our results and error analysis,

that emoticon-based DS is less suited for Arabic. Next, we further investigate this hypothesis by exploring learning rate for binary sentiment classifiers trained on two equally sized data-sets of Arabic and English tweets, where both were automatically annotated using positive/negative emoticons.

5.4.1.2 Learning Curves on Emoticon-based data-sets: Arabic vs. English

The purpose of this section is to explore how sentiment classifiers trained on similar sized data-sets that were annotated with the same sentiment markers will perform on different languages. Specifically, we investigate whether the use emoticons for automatically annotating tweets for sentiment is less suitable for Arabic as compared to English. For this, we assess the learning rate of sentiment classifiers trained on two emoticon-based data-sets (see page 55):

1. **Arabic data-set**, which is comprised of 1M tweets randomly sampled out of our Emo2 data-set.

2. **English data-set**, which is a random sample of 1M tweets out of the Emo-Eng data-set.

Both data-sets were built and automatically annotated for sentiment using emoticons and both contain a balanced class distribution, i.e. an equal number of positive and negative instances. Both models are trained using word-based n-grams (1g+2g) as features. The trained models are evaluated against manually annotated test-sets.[3]

The learning curves are displayed in diagrams 5.3 and 5.4. Each classifier is trained at several cutoff points, starting with using only 100k instances to train each classifier. The amount of training data increases on a 10%-basis in each run, until all one million instances are used for training. The x-axis in the diagrams indicates the size of training data used in each run.

[3]The Arabic SA classifier is evaluated using our test-set, while the English SA classifier is evaluated on a manually annotated test-set of 359 tweets that is created by Go et al. [81] and made publicly available. To account for the difference in size between the two test-sets, we ran follow-up experiments on a random sample of our test-set comprising only 436 tweets. On the reduced test-set, we observe an average improvement in performance at 2.48% as compared to our full test-set, with a top score at 54.82%.

Both curves reflect a general improvement in performance as more instances are used for training [31]. Also, both curves indicate SA classifiers experience declining performance at some points, which probably coincides with the addition of noisy training instances. Although the English SA classifier appears to be more vulnerable for noise at the beginning, the addition of more training data seems to benefit the classifier to overcome noise. The top scores with 1M tweets are at 53% for Arabic and 79.40% for English.

Overall, both SA classifiers (English and Arabic) benefit from adding more training data. However, the Arabic SA classifier is still attaining a lower performance on this task, which we attribute to noise emoticons introduce (see section 5.4.1.1).

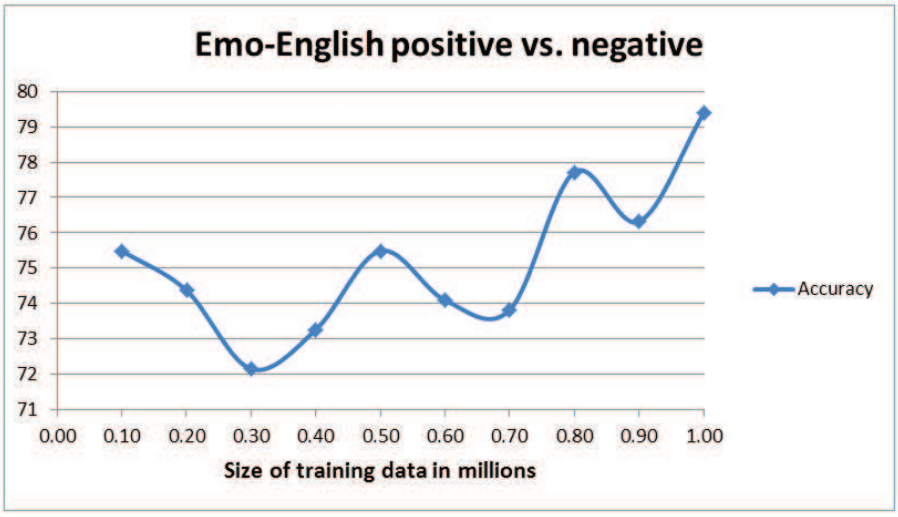


Figure 5.3: Learning curve on a 1M English emoticon-based data-set.

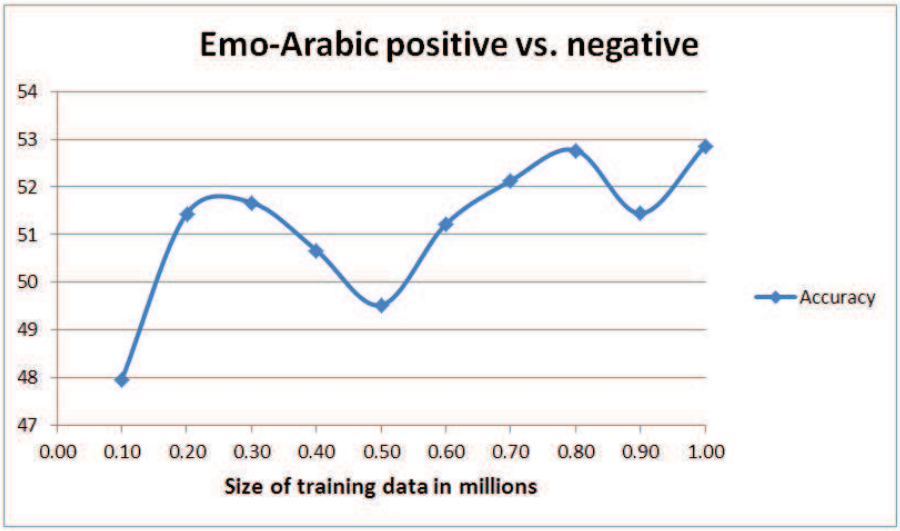


Figure 5.4: Learning curve on a 1M Arabic emoticon-based data-set.

5.4.2 Experiments on Extended Lexicon-based Data-sets

In this section, we experiment with extended lexicon-based data-sets (i.e. Lex-Pres2 and Lex-Aggreg2) that were built using lexicon-presence and lexicon-aggregation-based approaches to DS (see page 55). Again, the purpose here is to investigate the impact of lexicon-based training data-sets's expansion on the binary sentiment classification task. Results are summarised in table 5.8.

Lexicon-presence-based (Lex-Pres2) data-set: The best performance is recorded with the Twitter-specific features at 56.21%. Table 5.9 shows that the Twitter-specific features attained the lowest classification error rate in this set of experiments at 0.437 with a medium effect size at 0.34. The average performance with Lex-Pres2 is at 53.65%, which is 1.86% lower than LexPres1 on this task. Despite being 17.7 times larger than LexPres1, the results indicate that the the use of LexPres2 has not led to any improvement in the average accuracy compared to LexPres1. In contrary, the addition of more training data has been shown useful for the emoticon-based approach (section 5.4.1). This might suggest that the addition of more data can be more useful for some approaches (e.g. emoticon-based) than others (e.g. lexicon-presence-based).

Positive vs. Negative

	Lex-Pres2 Data-set		Lex-Aggreg2 Data-set	
	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.531	66.51	0.531	66.51
Stem n-grams	0.550	54.28	0.512	51.40
Stem n-grams + Morph	0.497	50.76*	0.448	47.23*
Stem n-grams + Affec-cues	0.542	53.96*	0.441	47.18*
Stem n-grams + Lang-style	0.552	54.01	0.534	52.50*
Stem n-grams + Twt-specific	**0.574**	**56.21***	**0.543**	**53.60***
Comb. of all feat.	0.523	52.68*	0.475	49.24*
Average	0.540	53.65	0.492	50.19

Table 5.8: Binary classification positive vs. negative on the lexicon-based data-sets. Underline denotes a statistically-significant difference vs. majority baseline ($p<0.05$). * denotes a statistically-significant difference vs. stem n-grams baseline ($p<0.05$).

Lexicon-aggregation-based (Lex-Aggreg2) data-set: The best accuracy score is achieved with the Twitter-specific feature-set, similar to Lex-Pres2. Table 5.9 indicates that this feature set is able to attain the lowest classification error at 0.464 with a large effect size of 0.58. The average accuracy score remains around 50.19%. The extended Lex-Aggreg2 data-set is 17.3 times larger than Lex-aggreg1 used in part one (page 133). Like the lexicon-presence method, the extension of the data-set has not yielded any improvement to the overall performance, as compared to the results achieved using Lex-Aggreg1. The average accuracy score at 50.19%, which is generally lower than that recorded on Emo2 and Hash, and even less than the Lex-Pres2's average accuracy score (further discussion in section 5.4.2.1).

5.4.2.1 Error Analysis for Lexicon-Based DS Data-set

Similar to the investigation conducted on the emoticon-based data-set (section 5.4.1.1), we manually labelled a random sample of 316 misclassified instances in order to investigate the accuracy of lexicon-based approach against human annotations.[4] The results reveal this approach to be more robust and less noisy, as the manual labelling (for positive and negative instances) has matched the automatic lexicon-based labelling in 62.03% of the cases, as compared to only 34.23% reached with the automatic emoticon-based labelling. The percentage of instances with

[4]We experimented with a 10:90 test:train split of the lexicon-based data-set in order to identify possible causes for errors within the automatically obtained sentiment labels.

	LexPres2			**LexAggreg2**		
	χ^2 ***(p-value)***	***Effect size***	***Class. error***	χ^2 ***(p-value)***	***Effect size***	***Class. error***
Stem n-grams	740.41 (0.000)	0.582	0.4571	1159.97 (0.000)	0.728	0.4860
Stem n-grams + Morph	1454 (0.000)	0.816	0.4924	1856.16 (0.000)	0.922	0.5277
Stem n-grams + Affec-cues	948.22 (0.000)	0.671	0.4603	2036.17 (0.000)	0.965	0.5281
Stem n-grams + Lang-style	435.19 (0.000)	0.446	0.4599	628.95 (0.000)	0.536	0.4750
Stem n-grams + Twt-specific	260.657 (0.000)	0.345	0.4379	747.83 (0.000)	0.584	0.4640
Comb. of all feat.	1222.58 (0.000)	0.748	0.4732	1680.76 (0.00)	0.877	0.5075

Table 5.9: Comparison of performance using different feature-sets on LexPres2 and LexAggreg2 data-sets.

automatic lexicon-based labels that contrasts with those assigned manually is at 20.25% compared to 36.63% with the emoticon-based data-sets (page 138). In addition, the manual annotation identified 10.13% of tweets as mixed (both positive and negative sentiments) and 3.16% as neutral. The rest of the instances were labelled as 'unclear sentiment orientation'. It is interesting to see that a random sample out of the lexicon-based data-set has more mixed tweets (10.13%), as compared to 5.94% of the emoticon-based data-set (page 138). This suggests more presence for mixed instances in an auto-labelled data using a lexicon-based method than that with emoticon-based. Consequently, the addition of more data has shown useful for emoticon-based classifiers, while adding more data has resulted in introducing more noise/confusion (i.e. mixed instances) to the classifiers trained on lexicon-based data [2].

Besides mixed instances, there are several potential reasons for classification error associated with a lexicon-based approach. Negation tokenisation and scope are problems we observed. For negation tokenisation, we observed that deficiencies in addressing negated words can occur when users omit the white-space word-boundary when using one of the most popular negators in Arabic, namely *la* (table 5.10). This is possibly because omitting white-space with this negator will not affect the readability of text, i.e. will not be perceived by the reader as a misspelling. However,

Negator	Negated word with white-space	Negated word without white-space
لَا (*not*)	لَا ينجح (*not successful*)	لَاينجح (*not* successful)

Table 5.10: Examples of negated words *with* and *without* omission of white-space word-boundary.

for a computer-based tokeniser, the task of identifying such a pattern is not trivial (table 5.10). The current version of MADAMIRA can identify/tokenise a negator as a particle only if a proper word boundary is used (e.g. white-space). A non-publicly available version of MADAMIRA uses morphological analysis to perform this type of tokenisation and, then, can correctly capture and tokenise negation even when word boundary is omitted. However, it uses components that require an LDC licence.[5] We are not aware of a publicly Arabic tokeniser that can address this issue by far.

Although we have accounted for negation in both lexicon-based methods using a negation-switch method (following Taboada et al. [165], see page 52), the current implementation for negation-scope handling is still far from optimal. That is, the negation scope we have considered (described in detail on page 52) is straightforward and tackles negated words within a short distance from negators (i.e. one or two tokens away), which have been identified as common scopes for negation [159]. However, it seems that a more fine-grained syntactic analysis is likely to enhance performance of sentiment classification. This includes, for instance, addressing negated words located at more variable distances, as in example #9, and/or accounting for more diverse forms of expressing negations [165], as in example #10.

9 لعب تشيلسي مَا فيه اي نوع من انوَاع الَاحترَافيه

Chelsea F.C.'s playing has nothing to do with professionalism.

10 مستحيل ان تكره اردوغَان

It is just impossible to hate Erdoğan.

Other sources of classification error we observed with this approach include limited coverage of the sentiment lexicon employed, such as dialectal sentiment-bearing words [60] and shorter/informal expressions used to convey emotions, which is also

[5](Nizar Habash, personal communication, April 18, 2016)

observed on English [163]. Below are examples of sentiment-bearing expressions that were not captured during the sentiment annotation process:

وصخ

A dialectal sentiment-bearing adjective meaning *dirty*.

وعّع

An informal expression typically used to convey a feeling of *disgust*.

اوف

An informal expression typically used to express a *growl*.

5.4.3 Summary of DS Experiments Part 2

Results of the Twitter's conventional markers and lexicon-based methods on the binary positive vs. negative task reveal that:

Conventional-markers-based methods:

- Extending the emoticon-based data from 121.5k instances in Emo1 to 1.2M instances in Emo2 (9.6 larger) has resulted in an average accuracy improvement of 2.56%. This indicates that more data is useful for the emoticon-based DS approach [31], but the improvement is generally slow. The best individual performance is attained with the morphological feature-set at 64.82%, which is 12% accuracy improvement over the stem n-grams baseline (table 5.5). This indicates the utility of the rich morphological features extracted using a publicly available tool not only with a gold-standard data (page 93), but also with an auto-labelled data using emoticons.

- As compared to English, our investigations in section 5.4.1.2 show that sentiment classifiers for English are able to to perform better than the Arabic ones with the same amount of training data obtained automatically using emoticons. Using 1M instances each, the Arabic sentiment classifier achieved 53.0% accuracy, while the English sentiment classifier attained 79.40% accuracy. Nevertheless, both classifiers show positive effect for adding more training data. A possible extension/improvement for this investigation is by evaluating the classifiers on the same test instances, i.e. original tweets and their gold-standard

translation [28]. That is because the test-sets currently used in section 5.4.1.2 are similar in size and both manually annotated, but ultimately each test-set has different instances. This might result in introducing bias in the results of one or both models.

- As for the hashtag-based data-set (Hash), the results show that this data-set is able to attain its best accuracy performance at 69.58%, which is 4.76% better than the best accuracy score achieved with Emo2 at 64.82%. This is interesting, noting that Emo2 is about 8.6 larger in size than Hash data-set. This suggests hashtags as less noisy, less ambiguous and more suitable for SA on Arabic tweets than emoticons, which confirms previous findings by AlMutawa [22]. An error analysis on the emoticon-based data-sets revealed that noise/ambiguity in this context involve the sarcastic use of emoticons [100] and mistakenly interchanging/mistyping of emoticons [22, 16] (section 5.4.1.1).

- Finally, combining emoticon- and hashtag-based data-sets has resulted in 6% improvement in average accuracy over Emo2 and Hash (individually) across all feature-sets. Nevertheless, the best individual accuracy score is still attained by Hash data-set at 69.58% using stem+affective-cues features (table 5.13).

Lexicon-based methods:

- The lexicon-presence-based method is superior to the lexicon-aggregation-based method (table 5.8). We hypothesise that this can be attributed to the considerable presence of mixed instances, i.e. tweets with positive and negative indicators, in the lexicon-aggregation-based data-set. Unlike the lexicon-presence-based method, wherein mixed instances are excluded simply because this method relies on the presence of positive or negative sentiment-bearing words for automatically determining a sentiment label, in the lexicon-aggregation data-sets they are kept in. This is because both positive/negative indicators contribute to the overall sentiment score in the lexicon-aggregation method. The sign of this summed up score (+ or -) for each tweet is then

used to assign it with a sentiment label (details on page 50). Consequently, the Lex-Aggreg2 data-set includes a total of 71,628 mixed instances (representing 14.69% of Lex-Aggreg2), the presence of which we anticipated at first might bring more diversity to this training data and therefore contribute in making the resulting classifiers more robust/accurate. However, the results suggest that mixed instances have resulted in more confusion to the sentiment classifiers.

- The average accuracy scores indicate that the extension of the lexicon-based data-sets from 78.5k in Lex-Pres1 to 471k in lex-Pres2 has not resulted in any improvement. In general, both lexicon-based approaches perform worse. That is, Lex-Pres1 (smaller data-set) is 1.86% accuracy better than Lex-Pres2 (larger data-set), while Lex-Aggreg1 (smaller data-set) is 0.62% accuracy better than Lex-Aggreg2 (larger data-set). It appears that the benefit of increasing the size of training data can vary across approaches. In addition, results of lexicon-based methods are worse than those reported on English, e.g. Zhang et al. [186] reported an accuracy score of 85.4% using a similar combined approach (Lexicon-based+ML) on English tweets. This might suggest that a combined approach of lexicon-based+ML is less effective in the context of SA of Arabic tweets (e.g. compared to hashtag-based DS).

- The coverage of the lexicon used is a key element in the effectiveness of the lexicon-based approaches [165]. As such, the application of lexicon-based approaches to a continuously evolving medium like Twitter is likely to be problematic and inflexible [186, 51]. To illustrate, existing opinion lexica might lack the presence of Twitter-specific characteristics like abbreviated/informal sentiment expressions (section 5.4.2.1) and spelling variations resulting in lower recall scores. For instance, we recorded recall scores at 0.543 for Lex-Pres2 and 0.514 for Lex-Aggreg2 (table 5.11). In this context, Zhang et al. [186] argue that keeping the sentiment lexicons up-to-date by manually adding the recently emerged sentiment expressions can be a hard task. Recent attempts have been made to address this problem for Arabic (both manually and semi-

supervised using rules/patterns) and maximise coverage of sentiment lexica, e.g. by including sentiment-bearing dialectal/slang expressions [65] and idioms [99]. In addition, El-Sahar and El-Beltagy [65] propose a rule-based system and hand-crafted patterns to extract trending/creative sentiment-bearing expressions, which have shown beneficial results on Arabic tweets. Others, like Salameh et al. [153], used an automated means to create sentiment lexica by employing popular algorithms, such as the Pointwise Mutual Information (PMI), which calculates a score for each given word that reflects its association with a positive/negative sentiment. Such methods can be employed for automatic expansion, i.e. improving coverage, of sentiment lexica utilised in lexicon-based methods for SA. Zhu et al. [187] have proven this approach to be useful for SA on English tweets.

Data-set	Precision	Recall	F-score
lexicon-presence Lex-Pres2 DS			
positive	0.400	0.728	0.516
negative	0.766	0.450	0.567
lexicon-aggregation Lex-Aggreg2 DS			
positive	0.389	0.788	0.521
negative	0.779	0.376	0.507

Table 5.11: Recall, precision and F-scores for lexicon-based DS methods: positive vs. negative (stem baseline).

- Finally, to assess the effectiveness of lexicon-based approaches, we validated both methods (a lexicon-based and a combined lexicon-based + ML-based) against human annotations (our independent test-set). The lexicon-based method simply utilises the presence of sentiment-bearing words to assign an instance with a sentiment label [165] (page 50). The combined approach 1) uses a lexicon-based method to obtain training instances that then 2) used to train a machine learning sentiment classifier [186]. Both the lexicon-based approach and the combined approach were used to automatically predict sentiment labels for our test-set. In order to conduct a fair comparison between methods, the instances classified as neutral by the lexicon-based method were excluded. Following that, a trained ML classifier was used to predict the sentiment labels

(positive or negative) for the remaining instances.[6] The lexicon-based method (on its own) reached an accuracy score of up to 65.8%, while the combined approach (lexicon-based + ML-based) attained an accuracy score of 53.6%, when both approaches evaluated against the same set of 838 instance. This suggests that, in the context of Arabic tweets, lexicon-based approach (on its own) is more effective than when combined with an ML-based method. A possible explanation is that the lexicon-based method relies only on our combined and manually created/filtered lexicon that is composed of context-independent sentiment-bearing entries to predict the sentiment label, following Taboada et al. [165]. In the ML approach, in contrast, an ML classifier will rely on context-independent (e.g. hate) and context-dependent/indirect-clues (e.g. Al-Asad) sentiment indicators that the model will infer [167, 143]. Thelwall et al. [167] reported an average accuracy of 62.65% using a lexicon-based method on English tweets, while studies on Arabic all targeted certain dialects and, in accordance, reported varying scores ranging between 59-80% [20, 9, 62, 171] (section 5.2). This variation can be attributed to reasons like: the varying degrees of difficulties in tackling different dialects, and differences in sizes of data-sets used. In this work, we explored goodness of SA systems on a multi-dialect design. Experimenting on multiple dialects seems to be more difficult, as the SA classifiers will be exposed to a wider diversity of lexical variation. However, it has the potential for developing general-purpose tools/corpora.

[6] Accuracy is calculated on a subset of 838 instances (38.4% of test-set) with instances with no sentiment-bearing words are excluded, following Taboada et al [165]. Further discussion on neutral/empty-text instances (with no direct sentiment indicators) is on page 15.

	Polar vs. Neutral					
	Data size	Best F *(feat)*	Best Acc. *(feat.)*	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Fully sup. (GS1+GS2)	8k	0.735 *(stem)*	73.99 *(stem)*	530.403 (0.000)	0.387	0.2603
Emoticon-based DS	121.6k	0.950 *(morph)*	95.19 *(morph)*	98.572 (0.000)	0.315	0.0511
Lexicon-presence DS **	78.5k	0.953 *(stem)*	95.36 *(stem)*	94.817 (0.000)	0.307	0.0480
Lexicon-Aggreg. DS	83.2k	0.910 *(stem)*	91.11 *(stem)*	137.329 (0.000)	0.370	0.0891

Table 5.12: Comparison of performance of a fully-supervised, emoticon-based, lexicon-presence-based and lexicon-aggregation-based approaches (stem n-grams) on the independent test-set with respect to accuracy. ** indicates the best performing approach on this task.

	Positive vs. Negative					
	Data size	Best F *(feat.)*	Best Acc. *(feat.)*	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Fully sup. (GS1+GS2)**	3.7k	0.780 *(twt-specific)*	77.97 *(morph)*	37.480 (0.000)	0.132	0.2377
Emoticon-based DS	1.2M	0.590 *(morph)*	64.82 *(morph)*	54.644 (0.000)	0.157	0.4722
Hashtag-based DS	130.2k	0.674 *(Affec.-cues)*	69.58 *(Affec.-cues)*	118.466 (0.000)	0.232	0.3078
Emoticon+ Hashtag-based DS	1.3M	19.752 (0.000)	0.621 *(stem)*	64.50 *(morph)*	0.095	0.3719
Lexicon-presence DS	415.8k	0.574 *(twt-specific)*	56.21 *(twt-specific)*	740.41 (0.000)	0.582	0.4571
Lexicon-Aggreg. DS	487.5k	0.543 *(twt-specific)*	53.60 *(twt-specific)*	1159.97 (0.000)	0.728	0.4860

Table 5.13: Comparison of performance of a fully-supervised, emoticon-based, lexicon-presence-based and lexicon-aggregation-based approaches (stem n-grams) on the independent test-set with respect to accuracy. ** indicates the best performing approach on this task.

	Positive vs. Negative vs. Neutral					
	Data size	Best F *(feat.)*	Best Acc. *(feat.)*	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Fully sup. (GS1+GS2)	8k	0.641 *(semantic)*	64.10 *(semantic)*	123.926 (0.000)	0.187	0.3660
Emoticon-based DS	121.5k	0.70 *(stem)*	69.67 *(stem)*	102.138 (0.000)	0.325	0.3032
Lexicon-presence DS **	78.5k	0.71 *(stem)*	71.57 *(stem)*	104.883 (0.000)	0.316	0.2842
Lexicon-Aggreg. DS	83.2k	0.630 *(stem)*	64.96 *(stem)*	176.261 (0.000)	0.427	0.3503

Table 5.14: Comparison of performance of a fully-supervised, emoticon-based, lexicon-presence-based and lexicon-aggregation-based approaches (stem n-grams) on the independent test-set with respect to accuracy. ** indicates the best performing approach on this task.

5.5 Discussion of DS Results

The experiments presented in this chapter investigate a number of DS approaches for automatically labelling larger data-sets for Arabic SA. This includes assessing the ability of DS approaches to outperform traditional fully-supervised machine learning approaches that are relying on manually-annotated data-sets.

DS on subjectivity and sentiment classification: The results seem to indicate that DS works well for subjectivity analysis, distinguishing neutral vs. polar instances (table 5.12). However, we anticipate that this is partially because the neutral class in training and test data is predominately in MSA, while the polar class is mostly in DAs (section 5.3.4). As such, it seems that the models mostly learn to distinguish MSA vs. DAs. In chapter 7, we re-assess the performance of these models in order to identify how well models will perform on tweets retrieved from the live Twitter stream (section 7.2).

In contrary to subjectivity analysis, DS proves to be difficult for SA, distinguishing positive vs. negative instances (table 5.13). The per-class F-scores indicate that for the emoticon-based DS, detecting positive instances seems to be problematic (see table 5.15), which we anticipated to be the case for positive class mainly due to

the amount of noise and ambiguity introduced when using emoticons as sentiment-labels, especially positive ones (see page 140). As for the lexicon-based DS methods, both approaches tend to maintain fairly balanced per-class F-scores for positive and negative.

Data-set	Precision	Recall	F-score
emoticon Emo2 DS			
positive	0.332	0.406	0.366
negative	0.663	0.589	0.624
hashtag Hash DS			
positive	0.560	0.376	0.450
negative	0.730	0.851	0.786
lexicon-presence Lex-Pres2 DS			
positive	0.400	0.728	0.516
negative	0.766	0.450	0.567
lexicon-aggregation Lex-Aggreg2 DS			
positive	0.389	0.788	0.521
negative	0.779	0.376	0.507

Table 5.15: Recall, precision and F-scores for DS methods: positive vs. negative (stem baseline).

In order to investigate possibilities for improving the performance of DS methods on sentiment classification task, we conducted a second round of investigations in which we explored the impact of expanding training data and using another conventional-marker of Twitter (i.e. hashtags):

Impact of extending training-set: Emo2 data-set is 9.6 times larger than Emo1 data-set. Results on Emo2 show an improvement of 2.56% over that attained on Emo1. The extension of the data-sets also included creation of a new hashtag-based data-set (Hash). Hash is composed of 130.2k instances and attained its best accuracy performance at 69.58%, which is 4.76% better than the best accuracy score achieved with Emo2 (1.2M instances) at 64.82%. When combined with Emo2, Emo2+Hash has improved the average accuracy score to 62.22%. This is a 6% improvement over the average accuracy score attained by Emo2 and Hash individually.

Lexicon-based data-sets (Lex-Pres2 and Lex-Aggreg2) are nearly 5 times larger than Lex-Pres1 and Lex-Aggreg2 but recorded no improvement. Instead, experi-

menting with larger data-sets slightly hurt the performance (section 5.4.3). This suggests that more training data using a lexicon-based method might result in introducing the classifiers to more noise (e.g. presence of mixed instances and use of positive words in negative context), especially with many dialectal and informal sentiment-bearing expressions being not captured with the currently available sentiment lexica for Arabic (section 5.4.2.1).

In this context, Banko and Brill [31] illustrate that the performance of learners in text classification tasks "can benefit significantly from much larger training sets". Our results suggest that exploiting considerably larger training data is beneficial for SA on Arabic tweets. However, this is not the case with all methods. For the SL and emoticon-based DS methods, on the one hand, the results show that the sentiment classifiers benefit from using a larger training data. The lexicon-based methods, on the other hand, using a larger training data slightly hurt the performance.

It is also interesting to note that the Hash data-set with only 130.2k instances is able to attain an accuracy score of 69.22% on the stem n-grams. This is 16.45% better than the score achieved by Emo2 with 1.2M instances. Thus, we conclude that better performance can be achieved with hashtags (even with much smaller training data) than emoticons on Arabic tweets (section 5.4.3). In this context, previous SA work on Arabic and English tweets observed hashtags to be more effective/reliable on their own [22, 109]. Others found that some conventional markers, i.e. emoticons or hashtags, are more suitable for the detection of some emotions than others. For instance, Purver and Battersby [135] observed that emoticons are better for detecting positive/happy, while hashtags are better in detecting negative/sad class. Kouloumpis et al. [109] found that both hashtags and emoticons are useful for SA on English tweets. However, this aspect can be language-dependent. An error analysis that we have conducted suggested emoticons in Arabic tweets to be highly noisy, especially for the positive class (see page 138).

5.5.1 Comparison with Previous Work

Emoticon-based and hashtag-based DS approaches: The use of emoticons as noisy labels has been shown to be successful, attaining accuracy scores of up to 83% for binary SA (positive vs. negative) on English tweets [81]. Using hashtag-based data-sets, previous work reported accuracy of up to 74% on English tweets [109]. As for Arabic, the best reported scores for emotion analysis (i.e. happiness and sadness) are at 51.35% using an emoticon-based data-set and 66.71% using a hashtag-based data-set [22]. We are not aware of previously published work that has addressed the use of emoticons to automatically label sentiment polarity (i.e. positive and negative) for Arabic tweets. The experiments presented in this chapter reached accuracy scores of up to 64.82% with the emoticon-based data-set and 69.58% with the hashtag-based data-set.

Lexicon-based SA approach: For English binary SA classification, using a lexicon-based approach, previous work reported up to 78.74% on reviews and 62.65% on tweets [165, 167]. All previous work on Arabic SA using lexicon-based methods has focused on one/two dialects with accuracy scores between 59-80% [20, 9, 62, 171]. In this chapter, we have investigated the utility of this approach on multi-dialectal data-sets. Our results reached an accuracy score of up to 65.8%, which is comparable to that reported on English tweets on the same task [167].

Combined: Lexicon-based + ML-based SA approach: The results presented in this chapter for utilising a combined approach, comprising a lexicon-based method (i.e. for obtaining training instances) + a machine-learning method (i.e. for training a sentiment classifier), reached up to 56.21%, which is worse than the accuracy reported on English tweets at 74% [186] and on single-dialect Arabic tweets at 79% [64].

Overall, it appears that a lexicon-based method can be more challenging on a non-factual-based text classification task like SA as compared to factual-based tasks, e.g. topic classification [128]. The reason is the difficulty of coming up with the right/optimal set of context-independent sentiment indicators, i.e. a sentiment

	χ^2 *(sig.)*	*Effect size*	*Class. error*
Lexicon-based	215.149 (0.000)	0.506	0.3424
Lexicon+ML-based	336.846 (0.000)	0.634	0.4642

Table 5.16: Comparison between Lexicon-based and lexicon + ML-based data-sets vs. gold-standard labels with respect to accuracy on a subset of 838 tweets of the independent test-set.

lexicon with sufficient coverage [128]. Furthermore, the issues associated with the tweets' genre (e.g. misspellings, spelling variation, slang – see page 12) contribute to the increasing difficulty in creating a sentiment lexicon with sufficient coverage [116]. To alleviate this problem, some studies [9, 171] have considered manually building sentiment lexica targeting specific dialects (e.g. Jordanian, Saudi and Egyptian), which proved beneficial. Others endeavour to enhance the coverage of the sentiment lexica by means such as creating lists of sentiment-bearing dialectal/slang expressions and idioms [65, 99].

Comparing different DS methods: In sum, DS approaches using conventional markers of Twitter and sentiment lexica presented in this chapter allowed assessing the performance of sentiment classifiers trained using different labelling techniques but intended to perform the same sentiment classification task [135]. Overall, the use of lexicon-based methods on a rapidly-changing medium like Twitter is prone to the coverage of sentiment lexicon used that can be directly influenced by issues like misspellings/spelling-variations typically encountered in tweets [51]. Our results indicate that the hashtag-based DS approach outperforms emoticon- and lexicon-based DS approaches for sentiment classification on Arabic tweets. Using hashtags-based distantly-labelled data in an ML-based approach can have the flexibility and scalability to better cope with the rapidly-changing nature of the Twitter stream.

5.5.2 Other Factors Influencing Performance in DS methods

Impact of normalising auto-labelling features: Removing features used for auto-labelling (i.e. emoticons and sentiment-bearing words) to avoid biasing the

data-set is likely to result in removing an important piece of information that would otherwise act as in informative clue for a sentiment classifier [81]. The problem is even more pronounced when there is a lack of independence of the auto-labelling feature from the accompanying text. Unlike cases wherein removing emoticons will not affect the overall sentiment orientation of a tweet, as in examples #11 to #14 (table 5.17), the sentiment orientation in other cases will be significantly affected by the removal of the accompanying emoticon, as in the case of sarcasm, see example #3 (page 140). Similarly, with the lexicon-based DS approach, normalising the sentiment-bearing words used to label the training examples can affect the ultimate emotion to be conveyed, as in examples #15 and #16 (table 5.17). That is, normalising the highlighted sentiment-bearing word will result in a seemingly neutral tweet. Ultimately, an essential element in DS approaches is based on the idea of whether the SA classifier will learn from the remaining features, without relying on the quality of labels [180]. In this context, Thelwall et al. [167] state that ML classifiers can infer sentiment from "indirect" sentiment indicators, e.g. words that are likely to appear in negative/heated political discussion, like *Israel*.

11	اليوم خسرت شخص احبه .. مَافيش حَاجه توصف احسَاسي :(*Today I've lost a beloved person .. nothing can describe my feelings :(*
12	سونيك درفت اسوَا لعبه في التَاريخ :(*Sonic Drift is the worst video game in the history :(*
13	صبَاح جميل لّلجميع :) *I wish a splendid morning for you all :)*
14	قصي خولي انت تمثل موهوب و كل مصر بتحبك :) *Kosai Khauli is a really talented actor that everybody loves :)*
15	صدي الملَاعب برنَامج مقرف *Sada Al-Malaeb is a **disgusting** TV show.*
16	موقف رَاءيع من جَامعه الدول العربيه تجَاه الغَارَات الَاسرَاءيليه عَلي غزه *A **wonderful** stance from League of Arab States towards the Israeli strikes on Gaza.*

Table 5.17: Examples of tweets automatically labelled for sentiments using DS methods.

5.6 Summary

One of the biggest challenges with Twitter data is its scalability (in terms of topic/lexical variation), as tweets cover almost every domain/topic and their language evolves over time [186, 112, 82]. This is likely to influence the coverage of training data-sets targeting such a domain. A possible remedy is the use of DS approaches, which uses readily available features like emoticons, as noisy labels in order to automatically annotate large amounts of data for learning topic/temporal-independent models. This approach has been shown to be successful for English SA, e.g. [81], and SA for less-resourced languages, such as Chinese [180] and Italian [34].

This chapter empirically evaluates the performance of existing DS approaches for SA on Arabic Twitter feeds. In addition, we conduct an error analysis to critically evaluate the results and give recommendations for future directions. We find that DS significantly outperforms fully supervised approaches for the binary task of subjectivity classification (polar vs. neutral) on our independent test-set, where we achieve 95.26% accuracy, which is a 22.97% improvement over previous fully super-

vised (GS1+GS2) results on this task. However, sentiment classification (positive vs. negative) proves to be difficult using DS approaches, with an average accuracy of 55.51% compared to the results of fully supervised at 75.91%.

A second round of experiments was conducted concentrating on the impact of increasing the size of emoticon-based and lexicon-based data-sets, and exploiting a newly collected data that uses another common conventional marker of Twitter, i.e. hashtags. The results indicate an improvement, with the best accuracy score attained with Hash data-set at 69.58%. This is a comparable score to the best accuracy attained with fully-supervised on this task (positive vs. negative) at 77.97% on the independent test-set, considering the fact Hash data-set was automatically labelled. Despite providing noisy labels, DS methods (e.g. hashtag-based) allow larger amounts of data to be rapidly and automatically annotated, and thus, can better cope with the topic shift issue observed in Twitter. That is, even with the best possible quality of gold-standard sentiment labels obtained manually, Twitter-based data-sets are vulnerable/susceptible to becoming less effective over time [61].

What next? In the following chapter, we explore the viability of a machine-translation (MT)-based approach that uses an off-the-shelf MT tool, and we assess how well this system will perform as opposed/compared to more resource-intense approaches, i.e. DS or fully-supervised approaches for SA on Arabic tweets. With MT-based approaches for SA, no annotated data-sets are needed, as MT-based methods rely on exploiting tools readily available for well-resourced languages, such as English.

Chapter 6

Machine Translation Based Approaches

The data-based approaches to sentiment analysis presented in chapters 4 and 5 rely on large, manually/auto-annotated data-sets or wide-coverage sentiment lexica, and, as such, might not be readily available in under-resourced languages. This chapter presents empirical evidence of an efficient tool-based SA approach that uses freely available machine translation (MT) systems to translate Arabic tweets to English, which we then label for sentiment using top performing publicly available English SA systems. Parts of this chapter are published in [142].

6.1 Related Work

In this section, we review previous attempts to utilise MT-based methods to transfer resources from/to English for SA across different domains and languages. Transferring resources from English can be used, e.g. to create training corpora, while transferring to English can be used, e.g. to employ English tools on translated text.[1]

MT on languages other than Arabic in non-Twitter domains.

Banea et al. [30] propose leveraging existing sentiment resources for English to be

[1] Tweets in particular have recently received attention with respect to the application of MT (e.g. the launching of the tweetMT-2015 workshop and shared-task; `http://komunitatea.elhuyar.org/tweetmt/`

transferred using MT into other languages. The aim is to overcome the issue of a lack of resources available for SA in less-studied languages. The study targeted Romanian as an under-resourced language. The authors translated a set of Romanian sentences taken from news into English (using a commercial MT system) and used an existing SA system for English to obtain labels for subjectivity classification (polar vs. neutral). The authors reported an F-score at 0.678 and deduced that MT is a viable alternative for the construction of resources/tools for SA in a new language.

In addition, Denecke [53] followed an MT-based approach for obtaining sentiment labels for a data-set of German movie reviews. The author translated the data-set into English using a commercial translation system. To assign sentiment labels (positive or negative), the author used an existing sentiment lexicon for English (SentiWordNet) to aggregate the overall sentiment orientation score of each data instance and assign it with a sentiment label (i.e. following a lexicon-based method). These sentiment labels were then evaluated against labels assigned based on a star-rating method (e.g. a review with 4 or 5 starts will automatically be positive and a review with 1 or 2 starts will be negative). Performing binary (positive vs. negative) classification, the authors reported an accuracy score of 66% and an F-score of 0.620.

Wan [170] proposes a co-training approach to tackle the lack of Chinese sentiment corpora by employing Google Translate as a publicly available MT service to translate a set of annotated English reviews into Chinese. The English reviews and their Chinese translations were used to train two SVM classifiers and then combined into a single sentiment classifier. The resultant classifier was then used to classify a held-out test-set of Chinese reviews and their English translation. Utilising word-based n-grams as features, the author reported accuracy scores at 77.1% on the Chinese SA classifier, 76.9% on the English SA classifier and 81.3% on the combined classifier (Chinese and English) for binary (positive vs negative) classification.

Duh et al. [59] experimented on a data-set of Japanese, French and German reviews. The data was translated into English using Google Translate. Training SVM classifiers using word n-gram features, the authors reported the best performance (positive vs. negative) on held-out data with the German data-set at an accuracy

score of 77.0%, followed by French at 75.6%, and 69.4% on Japanese. The authors noted that language mismatch can play a role, i.e. German and English are both Germanic languages and can have a better degree of overlapping with each other than Japanese, as an Altaic language. In section 6.2.2, we show that the results obtained on Arabic-English are close to that obtained on Japanese-English by Duh et al. [59].

MT on languages other than Arabic in Twitter data.

As for tweets, Agarwal et al. [12] examine SA on a data-set of foreign tweets[2] that was translated into English using Google Translate. The data-set was then manually annotated into positive, negative or neutral, excluding tweets that were found hard to understand by human annotators (i.e. not well translated). The resultant data-set is balanced and used to train SVM classifiers with a 5-fold CV setting. The authors reported an accuracy score of up to 75.39% for binary (positive vs. negative) and 60.50% for three-way (positive vs. negative vs. neutral) classification by combining word n-grams and semantic features. Our work differs from the work of Agarwal et al. [12] in utilising an MT-based method for obtaining sentiment labels without involving human annotators, i.e. to avoid the cost of obtaining sentiment labels manually.

Balahur and Turchi [28] investigate the use of an MT system (Google) to translate an annotated corpus of English tweets (SemEval's manually annotated data [123]) into four European languages (Italian, Spanish, German and French). The purpose is to obtain annotated training-sets for learning five sentiment classifiers. In addition to English, they train an SVM classifier for each of the target languages using word n-gram features. Finally, the SA classifiers were evaluated on a held-out test-set, which was also translated into the same target languages (translation of test data was manually corrected for each of the target languages). For three-way classification (positive vs. negative vs. neutral), the authors reported an accuracy score of 64.75% on the English held-out test-set. For the other languages, they reported accuracy scores ranging between 60 - 62%. Hence, they conclude that it is possible to obtain

[2]The authors did not specify what languages.

high quality training data using MT, which is an encouraging result to motivate our approach.

MT on Arabic in non-Twitter domains.

As for Arabic, Bautin et al. [36] investigate MT to aggregate sentiment from multiple news documents written in nine different languages, including Arabic. The collected text was then translated into English using a web translator. Then, they used a lexicon-based approach (similar to that described on page 27) to assign sentiment labels (positive or negative) for each data instance. The authors argue that despite the difficulties associated with MT (e.g. information loss), the translated text still maintains a sufficient level of captured sentiments for their purposes. This work differs from our work in terms of domain and in measuring/evaluating summary consistency (i.e. the polarity correlation across different languages) rather than SA accuracy.

Rushdi-Saleh et al. [149] present an opinion corpus for Arabic comprising movie reviews (OCA) and its English translation (EVOCA). The data-set is balanced (number of positive and negative examples is equal) and the English translation was obtained using a free web translator. Similar to Denecke [53], the authors utilise the ratings associated with reviews to automatically determine their sentiment labels (positive or negative). They train SVMs on OCA and EVOCA using word-based n-grams and 10-fold CV setting for binary (positive vs. negative) classification. The authors report an F-score of up to 0.90 on the original Arabic text (OCA), and 0.88 on its English translation (EVOCA). Our work is different from this work with respect to domain and in exploiting an SA system that was originally trained on English text instead of training an SA system on a translated text (section 6.2.2).

MT on Arabic in Twitter data.

As for Arabic tweets, a recent study by Salameh et al. [153] on a data-set of 2k Syrian tweets has looked into the impact of MT on sentiments. First, the authors manually annotated the Syrian tweets as positive, negative and neutral. Second, they translated the tweets using an in-house MT system that was trained on Arabic

news data. Third, the translated data was then annotated for sentiment automatically and manually. To automatically obtain sentiment labels for the translated tweets, they employ an SA system with an SVM trained on English tweets that they previously created using SemEval's data-sets [187]. To obtain the manual sentiment labels, the authors asked human annotators to label the translated tweets. Fourth, both the automatically produced and manually assigned sentiment labels were evaluated by matching them against human annotations of the original Arabic tweets. On three-way classification (positive vs. negative vs. neutral), they reported an accuracy score of 78.11% with the automatically generated sentiment labels (generated from SVM) and 71.05% with the manually obtained sentiment labels (manual annotations of translated data). Subsequently, the authors deduce that translation errors can be misleading for human annotators, but do not seem to have the same impact on SA systems that automatically predict labels (i.e. using a machine learning classifier). They attribute this phenomena to the ability of the ML-based SA system to learn the correct sentiment labels, provided that translation errors occur systematically. This interesting finding, that ML classifiers tend to perform better than human annotators on auto-translated text, is encouraging for our work presented in this chapter. Our work differs from the work of Salameh et al. [153] in, first, experimenting with a larger and multi-dialect Arabic Twitter data-set (our independent test-set of >3.5k tweets). Second, we avoid manual annotation by employing a publicly available SA system for English (i.e. the Stanford Sentiment Classifier by Socher et al. [160]) to assign sentiment labels for the English translation of our Twitter data. Third, unlike the in-house MT system used by Salameh et al. [153], we utilise publicly accessible MT tools (e.g. Google) to assess how well an MT-based SA system will perform with such off-the-shelf MT systems.

Summary of related work: Overall, there is a fair amount of prior work on leveraging English data/models to improve sentiment analysis in other languages. For that, previous work described in this section has adopted two main scenarios: 1) translating English corpora into another language, projection of sentiment labels, and training SA models on the translated data; or 2) translating unannotated data

from a less-resourced language into English, employing an existing English SA system to obtain sentiment labels, and projection of obtained sentiment labels back into the source language. To the best of our knowledge, no studies have considered exploring the impact of an MT-based method for SA in Arabic tweets prior to our work published in [142].

6.2 Approach

In this chapter, we follow the scenario in which we assume that we have Arabic tweets with no sentiment annotations and we employ an off-the-shelf MT system to translate them into English. Finally, the translated data is passed through an existing SA system for English to cheaply assign tweets with sentiment labels, i.e. avoiding the time and cost of manually obtaining these labels (figure 6.1). To the best of our knowledge, this study is amongst the first attempts to assess the impact of automatically translated data on the accuracy of SA of Arabic tweets.

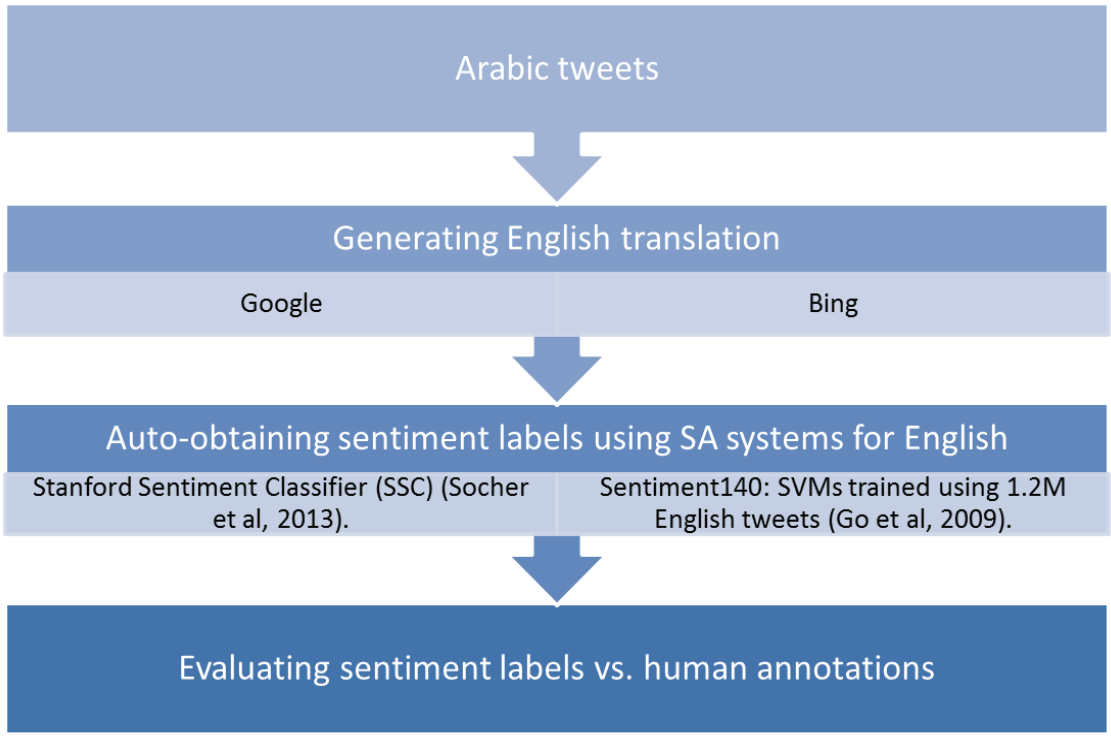


Figure 6.1: Architecture of an MT-based SA system.

6.2.1 Generating English Translation

In order to obtain English translation for our Arabic Twitter data-set, we employ two common and publicly available MT systems: Google and Bing translator services, following Balahur and Turchi [28] and Duh et al. [59]. Specifically, we translate our independent test-set (section 3.1.4 on page 54) to English, which we use to

evaluate SA systems/approaches through this work. Following Denecke [53], we do not perform any correction on the translated text, aiming to assess how well the SA systems will perform on translated (expected to be noisy) data.

6.2.2 Experiments on MT-based Approach: Using the Stanford Sentiment Classifier

This section presents an empirical evaluation of automatic sentiment annotations assigned by an English SA system on Arabic tweets translated to English using Google and Bing translators. In addition, we empirically benchmark the performance of this MT-based approach towards previous SA approaches (reported in chapters 4 and 5), including fully-supervised and distant supervision SA methods.

6.2.2.1 Sentiment Annotation

We use the Stanford Sentiment Classifier (SSC) developed by Socher et al. [160] to automatically assign sentiment labels to Arabic tweets translated into English. The choice of SSC is motivated by its superior performance (up to 85.4% accuracy) and the availability of its trained models for public [160].[3] That is, SSC is based on a deep learning (DL) approach, using recursive neural models to capture syntactic dependencies and compositionality of sentiments. SSC is trained on a data-set of English movie reviews. In particular, the authors utilise the Stanford Sentiment Treebank (SST) that includes >200k manually labelled phrases extracted from 12k reviews to capture meanings/sentiments of phrases of variable length. As such, the authors state that SSC is potentially useful for capturing sentiment from short pieces of text like tweets. Using a held-out test-set of 2210 reviews, Socher et al. [160] show that this model significantly outperforms standard models, such as NB and SVM, with an accuracy score of up to 85.4% for binary classification (positive vs. negative) at sentence level using word-based features.

We use SSC to to automatically assign sentiment labels (positive, negative and neutral)[4] to the translation of our independent test-set. Using Socher et al. [160]'s

[3] `http://nlp.stanford.edu/sentiment/.`

[4] SSC distinguishes between 5 sentiments, including very-positive, positive, neutral, negative,

approach for directly training a sentiment classifier will require a larger training data-set, which is not available yet for Arabic.[5]

6.2.2.2 Experiment Results

Following previous work (section 6.1), and since SSC was trained on word-based features, we report the MT-based method experimental results using only word-based n-grams features. Therefore, comparisons with previous approaches in this chapter are all performed using the word-based n-grams features only. Results are summarised in table 6.1. In addition, table 6.2 shows comparisons between the MT-based method (both with Google and Bing translators) in terms of accuracy against the manually assigned gold-standard labels of our test-set. Table 6.3 compares MT-based approaches with the previously best performing approach in our experiments for the different classification tasks.

Binary classification: Polar vs. Neutral. The combination of Bing translator and SSC (Bing+SSC) attains an accuracy performance of 63.10%, which is 6.46% significantly better than that achieved with Google translator+SSC on this task. Table 6.2 shows that Bing+SSC is able to attain a lower classification error rate on this task at 0.369 than Google+SSC with a medium effect size at 0.43. Overall,, the score achieved by Bing+SSC is still below the best score recorded in our experiments by far on this task at 95.36% by the lexicon-presence-based DS method (table 6.3). This is also lower than the performance attained with a fully-supervised method (GS1+GS2 data-set) on this task at an accuracy score of 73.99%. In section 6.2.2.3, we conduct an error analysis to better understand the underlying sources of error with the MT-based SA approach.

Binary classification: Positive vs. Negative. Again, the sentiment labels obtained by SSC on Bing translation (Bing+SSC) has reached a better accuracy score of 66.42%, which is 6.41% better than the accuracy recorded with Google+SSC

and very-negative. For the purpose of consistency with our previous experiments, all very-positive and very-negative were mapped to the standard positive and negative classes. In total, instances automatically classified as very-positive and very-negative account for <1% of our test-set.

[5]SSC was trained using a set of 215,154 unique, manually labelled phrases.

on this task (table 6.2). This accuracy score is slightly below majority baseline, but can be further boosted up to 76.24% when excluding positive/negative instances that were assigned a neutral label by SSC,[6] following Taboada et al. [165]. However, to make a fair comparison with previous approaches, table 6.1 displays the results attained on the entire test-set. As such, the top accuracy score with MT-based method is at 66.42% accuracy, which is 9.81% below the best recorded score on this task that is attained by a fully-supervised method (GS1+GS2 data-set) at 76.23% on word-based n-gram features (table 6.3). Nevertheless, it is interesting to note that the accuracy attained by Bing+SSC at 66.42% is close to the performance of the best performing DS-based method (hashtag-based data-set) at an accuracy score of 69.22% on word n-grams (page 138). This probably suggests that MT-based methods has more potential to be used with sentiment classification (positive vs. negative) than subjectivity classification (polar vs. neutral).

Three-way classification: Positive vs. Negative vs. Neutral. The best performance here is also attained by Bing+SSC at an accuracy score of 50.32%, which is 4.02% significantly better than Google+SSC. Table 6.2 shows that Bing+SSC on this task is able to attain a lower classification error score than Google+SSC at 0.496 with a large effect size of 0.74. The best score on this task in our previous experiments is attained by a lexicon-based-presence DS method at 71.57% on word n-grams, which is notably better than the score of Bing+SSC (table 6.3). In section 6.2.2.3, we investigate possible reasons for the superiority of Bing translation over Google for SA in Arabic tweets.

Comparison with previous studies: It is worth mentioning here that Salameh et al. [153] reported better results with an accuracy score of up to 78.11% on the three-way classification task using an SA system that predicted sentiment labels on Arabic tweets translated to English (section 6.1). Since the authors use a different test-set, our results are not directly comparable. Our work also differs from the work of Salameh et al. [153] in using a larger (>3.5k tweets) and multi-dialectal Twitter

[6]The total number of positive/negative instances that were classified as neutral in this task is 507 tweets (23.22% of our test-set).

MT-based method on SSC

	Polar vs. Neutral		**Positive vs. Negative**		**Positive vs. Negative vs. Neutral**	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (b-mjr)	0.471	61.70	0.531	**66.51**	0.239	41.04
Google Trans.+SSC	0.505	56.64	0.509	60.01	0.420	46.30
Bing Trans.+SSC	**0.558**	**63.10**	**0.553**	66.42	**0.450**	**50.32**

Table 6.1: Binary and three-way classification on the independent test-set.

data-set, while the authors experimented on 2k Syrian tweets. Our previous review of literature revealed that better results can be reached with SA systems designed with a particular dialect in mind (section 4.1 on page 82). Another difference from Salameh et al. [153]'s work is that we use publicly available MT systems (i.e. Google and Bing) with no further corrections on translated data. As such, both MT systems (Google and Bing) we use will merely transcribe out-of-vocabulary (OOV),[7] see examples 3 and 4 in table 6.4. Salameh et al. [153], in contrast, used an in-house MT system that normalises all OOV words in translated text by replacing them with place-holders, which is expected to have a positive impact on alleviating/eliminating noisy features.

Summary: In sum, we observe the following:

- For subjectivity classification (polar vs. neutral), the MT-based SA classifier is able to attain a reasonable accuracy score of up to 63.10%, outperforming a majority baseline. Although this result does not compete with our best results on this task attained by a resource-intense and data-based DS approach, this is still close the results reported in previous work using MT approaches on Twitter data, ranging between 60-65% [28] (section 6.1).

- For binary sentiment classification (positive vs. negative), our MT-based SA system using Bing+SSC reaches a comparable performance at 66.42% to that achieved by a more resource-intense DS system, namely the hashtag-based

[7]OOV are terms encountered in input which are not present in a system's dictionary of known words [84].

	Polar vs. Neutral			**Positive vs. Negative**			**Positive vs. Negative vs. Neutral**		
	χ^2 ***(p-value)***	***Effect size (sig.)***	***Class. error***	χ^2 ***(sig.)***	***Effect size***	***Class. error***	χ^2 ***(p-value)***	***Effect size (p-value)***	***Class. error***
Google Trans.+SSC	323.42 (0.000)	0.302	0.4336	103.19 (0.000)	0.218	0.3999	1340.2 (0.000)	0.615	0.5370
Bing Trans.+SSC	662.08 (0.000)	0.432	0.3691	153.28 (0.000)	0.264	0.3357	1956.6 (0.000)	0.743	0.4968
Google Trans.+SSC vs. Bing Trans.+SSC	99.264 (0.000)	0.168	0.4013	6.634 (0.010)	0.056	0.3678	105.32 (0.000)	0.173	0.5169

Table 6.2: Comparison between Google and Bing translated data-sets with respect to accuracy (stem n-grams) on the independent test-set.

Task	**Previously best performing SA**			**MT vs. Best performing SA**	
	Method	F	Acc.	χ^2 ***(sig.)***	***Effect size (sig.)***
Polar vs. Neutral	DS (Lexicon-presence)	0.953	95.36	392.89 (0.000)	0.333 (0.000)
Positive vs. Negative	fully-sup. (GS1+GS2)	0.767	76.23	10.71 (0.001)	0.070 (0.001)
Positive vs. Negative vs. Neutral	DS (Lexicon-presence)	0.71	71.57	143.99 (0.000)	0.202 (0.000)

Table 6.3: Comparison between MT-based method (Bing + SSC) and best performing SA systems with respect to accuracy (stem n-grams) on the independent test-set.

DS approach at 69.22% on word n-grams. Thus, it seems that an MT-based SA method, exploiting publicly available tools, can provide another cheap, effective and fast way for obtaining sentiment annotation for Arabic tweets. Unlike the data-based methods (e.g. hashtag-based DS), sentiment labels in tool-based methods (e.g. MT-based) can be automatically obtained without training a new classifier. The fully-supervised system is still the top performing on this task at 76.23% on word n-grams, but the results suggest that MT-based SA can provide a cheap alternative with reasonable performance when no labelled data is readily available.

- The performance with the three-way classification is at 50.32%. A closer look at the results reveal the neutral class to have a very low F-score at 0.243. Socher et al. [160] ignored the neutral class for reporting results and focused only on positive and negative instances. Therefore, the model's ability to predict the neutral class on the original labels (manually assigned to the English instances) is not clear. However, a possible explanation is that the neutral class accounts for 19% of the data used to train SSC, making it a minority class and prone to be misclassified.

- It appears that Microsoft Bing MT is generally performing better than Google MT for translating Arabic tweets into English. At least, its impact/use with the SSC tool is better (we elaborate on this issue in section 6.2.2.3).

6.2.2.3 Error Analysis

The above results highlight the potential of an MT-based approach using publicly available tools (e.g. Bing and SSC) as a fast and cheap alternative for languages that lack a large training data-set annotated for SA, such as Arabic. In the following, we conduct a detailed error analysis to fully understand the strengths and weaknesses of this approach.

Google vs. Bing on Arabic tweets. First, we investigate the superior performance of Bing over Google MT by manually examining examples where Bing trans-

lated data is assigned the correct SA label (correct in this context implies matching labels assigned by human annotators), but Google translated data is assigned an incorrect SA label. We found that this is the case for 11.53% of instances of our test-set. We manually examined a random sample of 108 instances. This analysis reveals that one difference is the ability of Bing translator to maintain a better sentence structure while SSC uses neural networks to capture syntactic dependencies [160]. For instance, examples 1 and 2 (see Table 6.4) show cases wherein both translators are able to correctly translate sentiment-bearing words (e.g. love, traitors and killers), but Bing seems to be able to produce more meaningful (i.e. closer to human-produced) translations compared to Google. This has possibly given Bing an advantage in this context, especially as SSC is trained on a manually annotated English data-set. Bing is also found to perform better than Google for Hindi-English translation [55]. Both tools use a statistical machine translation mechanism that allows statistical models to learn from large amounts of parallel-corpora. However, detailed comparison for the underlying features of each tool is beyond the scope of this work.

Another difference that we observe between Bing and Google is shown in examples 3 and 4. These examples show cases where Bing translator is able to correctly capture and translate slightly misspelled sentiment-bearing words in the original Arabic text that Google failed to capture. This results in information loss as crucial sentiment-bearing words are transcribed. Capturing slightly misspelled words (e.g. words with two repeated letters) can give an advantage for Bing in this case as sentiment-bearing words in social media platforms are likely to be stressed by authors, i.e. expressive lengthening (see page 56).

SA on auto-translated data. In the following, we conduct an example-based error analysis of the MT-based approach for SA, aiming to find out the main sources of error when adopting this approach for Arabic tweets. Because results indicate a better performance with Bing translator, we therefore only consider examples where the MT-based SA system using Bing leads to a different SA label than that assigned by humans. For this, we inspect a random sample of 100 misclassified tweets. We

1

Example Tweet	ابحث عن الحب في اغَانِي فيروز
Google Trans.	Look for the love songs in turquoise.
Bing Trans.	Search the love in songs of Fairuz.
Human Trans.	Look for love in Fairuz's songs – referring to a famous singer.

2

Example Tweet	يسقط كل خونه الَاخوَان المسلمين القتله
Google Trans.	Each fall traitors Muslim Brotherhood in Egypt killers.
Bing Trans.	Down with all the traitors of the Muslim Brotherhood in Egypt the killers.
Human Trans.	Down with all the traitors and killers of the Muslim Brotherhood in Egypt.

3

Example Tweet	يقَال ان اَالكلب بشَار يريد عمل تفجير في الحج
Google Trans.	It is said that the Aalkalp Bashar wants to work in the bombing of the Hajj.
Bing Trans.	The **dog** reportedly Bashar wants a bombing in Hajj.
Human Trans.	That **dog** Bashar Al-Assad wants to bomb Hajj – referring to the largest annual religious gathering for Muslims.

4

Example Tweet	تويتر مَااقدر اوصف شنَاعّته
Google Trans.	I really appreciate what Twitter Describe the Hnaath.
Bing Trans.	Twitter what I describe his ugliness.
Human Trans.	I cannot describe how ugly Twitter is.

Table 6.4: Example tweets along with their Google, Bing and human translations (transcribed/not-translated words are underlined).

observe the following cases of incorrectly classified tweets (see Table 6.5):

	Example Tweet	Human Translation	Auto Translation	Manual label	Auto Label
1	ولي عهد بريطانيا طالع كشخه في الزي السعودي	Crown Prince of Britain looks very elegant in the Saudi attire.	Crown Prince of Britain climber Kchkh in Saudi outfit.	positive	negative
2	صبَاح الفل يَا مصر	Good morning Egypt.	Ehikioya o Egypt.	polar	neutral
3	وعشَان انكم معَايَا انَا امتليت حيَاه، امتليت حب	Because you are with me, I'm full of life and love.	And Ashan you having I Amtlat Amtlat love life.	positive	negative
4	هذَا الشبل من ذَاك الَاسد، الله يعَافيك و يطول بعمرك	A chip off the old block, God bless you with a healthy and long life.	That cub is from that lion God heal and go on your age.	positive	negative
5	فرّرحّه محمد بَالهدف	Muhammad's happiness with scoring a goal.	Frrhahah Muhammad goal.	positive	negative
6	يَا الله امطر اهل سوريَا بَالَامن وَالرزق	Oh God, shower people of Syria with safety and livelihood.	Oh God rained folks Syria security and livelihood.	positive	negative
7	القمه الحكوميه في دبي بصرَاحه عمل يستحق التقدير، روعه	Frankly, the Government Summit in Dubai is a splendid work that deserves recognition.	Government summit in Dubai Frankly work deserves recognition, splendor.	positive	negative

Table 6.5: Examples of misclassified tweets (transcribed/not-translated words are underlined).

- Examples 1 and 2 fail to translate the sentiment-bearing dialectical words, 'elegant' and 'Good morning', transcribing them as 'Kchkh' and 'Ehikioya' but not translating them. Example 3 represents a case of correctly translated sentiment-bearing words (love, life), but failed to translate surrounding dialectal text ('Ashan' and 'Amtlat'). Bautin et al. [36] point out that this type of contextual information loss is one of the main challenges of MT-based SA. Overall, MT systems on DAs are still performing less effectively as compared to MSA, mainly due to the lack of linguistic resources required, e.g. parallel corpora (see also see page 23) [185, 84]. As a result, one of the major challenges of an MT-based SA approach seems to be the use of DAs in social media platforms, such as Twitter. To investigate this issue, we study the correlation between language class (i.e. MSA or DA) and SA accuracy. The results indi-

cate a significant correlation (Pearson's correlation coefficient, $p<0.05$), with MSA outperforming DAs (figure 6.2). This confirms our hypothesis that DA is a major source of error for MT-based SA.

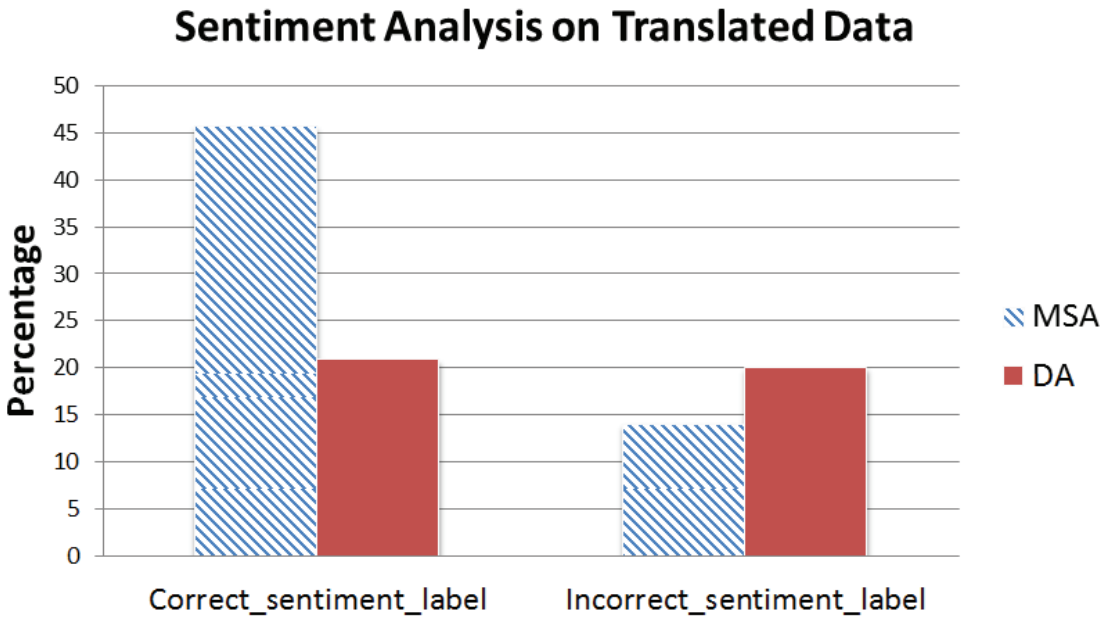


Figure 6.2: Performance of the MT-based sentiment classifier with respect to language class (MSA or DA).

- Failing to capture sentiment-bearing phrases/idioms,[8] see e.g. *that cub is from that lion* in example 4 (table 6.5), which can mean/correspond to *chip off the old block* or *like father like son* and is typically used in a positive context. Unlike the case in example 1, in this case the MT system correctly translated the individual words of an idiom/phrase. However, the resultant translation might not deliver the same sentiment orientation intended with that idiom in its source language. A possible remedy is by utilising existing idiom sentiment-bearing lexica (e.g. a recently created lexicon of idioms by Ibrahim et al. [99]) in which each entry (idiom/phrase) is assigned with a sentiment label (positive or negative). Such lexica can be used in a pre-processing stage to map identified entries in a given tweets to a single word that expresses the sentiment orientation assigned in the lexica (e.g. positive).

[8]Ibrahim et al. [99] define idioms as an expression that might not be understood from the individual meanings of its elements and can yield different sentiments when treated as separate words.

- Misspelled and, hence, incorrectly translated sentiment-bearing words in the original text – see example 5 'Frrhahah' ('happpinesss') with multiple repeated letters. The problem of misspelling is also highlighted by Abbasi et al. [2] as one of the challenges facing SA for Twitter data. Translation errors and/or missing translations as a result of (e.g. misspelling) have been identified amongst the main sources of error for MT-based SA systems [53, 28]. To reduce the impact of translation errors, some studies performed manual correction for translations, e.g. [28], while others opted to eliminate examples that were not well-translated as identified by human judges, e.g. [12] or replace OOV words with place-holders, e.g. [153]. As a further improvement for our investigations reported in this chapter in which we applied no correction to auto-translated data, we find the last solution (i.e. replacing OOV with place-holders) is in line with our goals of avoiding manual efforts for inspecting/correcting mistranslated instances. To identify OOV, a recent toolkit called *REMOOV* is developed for Arabic and made freely available for research community [56].

- Example 6 shows a correctly translated tweet, but with an incorrect sentiment label. We assume that this is a case of cultural differences: the phrase "oh God" can have a negative connotation in English [162]. Note that the Stanford Sentiment Classifier makes use of a manually labelled English sentiment phrase-based data, which may introduce a cultural bias. Salameh et al. [153] reported that tweets containing the automatic translation of "oh God" were manually annotated by English speakers annotators as negative, even though the sentiment labels obtained on the original Arabic tweets by Arabic speakers were positive.

- Example 7 represents a case of a correctly translated tweet, but with an incorrectly assigned sentiment label. We assume that this is due to changes in sentence-structure/word-ordering typically introduced by MT systems/tools [84]. In MT, re-ordering the translated words in a way that makes sense in the target language is known as *word-alignment/syntactic-ordering* [91, 84].

Balahur and Turchi [28] state that word ordering is one of the most prominent causes of SA misclassification of translated text. Note that SSC pays particular attention to sentence structure due to its "deep" architecture that might add to the model the feature of being sensitive to "compositional effects of sentiment", i.e. word ordering and overall meaningfulness [160]. Socher et al. [160] argue for the crucial role of the word ordering in such a semantic task like SA. In the following, we will verify this by comparing these results to another high performing English SA system that uses trained SVMs (section 6.2.3).

Conclusion: In sum, it seems that amongst the major challenges of an MT-based SA approach for Arabic tweets is the failure to translate a text or part of it, mainly because of misspelling and the use of DAs. Issues like dialectal variation and lack of standard orthography for DAs still present a challenge to MT [185]. This is especially true for tweets as they tend to be less formal resulting in issues like misspelling and individual spelling variations. However, with more resources being released for informal Arabic and Arabic dialects (further details on page 23), we assume that off-the-shelf MT systems/tools will improve their performance in the near future. In this context, a recent effort by the Advanced Technology Lab in Cairo has released a toolkit with multiple capabilities that include converting dialectal (Egyptian) Arabic to MSA.[9] The toolkit has been integrated to Bing, among other Microsoft products. The release of such a tool is anticipated to facilitate expanding Arabic NLP research by alleviating the noise caused by DAs and improve the quality of the extracted features.

6.2.3 Experiments on MT-based Approach: Using the Emoticon English Data-set (Emo-Eng)

This section investigates the viability of an MT-based SA system that pays less attention to grammatical structure (section 6.2.2.3), as compared to SSC. In particular, we employ another high performing publicly available English SA system that

[9] http://research.microsoft.com/apps/mobile/ShowPage.aspx?page=/en-us/projects/colloquial/

uses a large emoticon-based English Twitter data-set to train ML classifiers coupled with word n-grams [81]. Then, the trained model is used to assign sentiment labels for the English translation of our test-test (figure 6.1). Because results with Bing's translation of the test-set were generally better in our previous experiments (section 6.2.2), we report here on experiments using only Bing translator.

6.2.3.1 Sentiment Annotation

To obtain sentiment labels, we use the Emoticon-English (Emo-Eng) training data-set (page 47) to re-construct the SA system proposed by Go et al. [81]. The training data-set is composed of 1.6M tweets automatically labelled for sentiment based on the presence of emoticons. The data-set is balanced (number of positive and negative tweets is equal) and made publicly-available (page 47). Go et al. [81] used this data-set to train three ML classifiers (NB, MaxEnt and SVM).[10] They reported the best accuracy score using word n-grams features at 83.0% on a held-out test-set of 359 manually annotated tweets.

The choice of Go et al. [81]'s English SA system is motivated by the high accuracy reported on English tweets [81] and the freely-available training data. In addition, the system was re-evaluated in a recent study against a larger and more diverse benchmark English Twitter test-set and reported amongst the top performing systems [2]. Furthermore, Abbasi et al. [2] identified Go et al. [81]'s system performance as the most balanced/consistent across various test-sets among 15 other assessed systems, attaining an average accuracy score of 66.46%.

We use the Emo-Eng data-set and follow the steps described in [81] to train an SVM classifier using word n-grams as features. The trained model was then used to automatically assign sentiment labels (positive or negative) to the English translation of our test-set. We call this system GoEmo from now on.

[10]For the sake of consistency with all work reported in this thesis, we experiment with only an SVM classifier.

6.2.3.2 Experiment Results

Because the Emo-Eng training data-set of Go et al. [81] includes only positive and negative tweets, the experiments in this section will focus on the binary SA classification of positive vs. negative and compare the results against those obtained with Bing+SSC on this task.

Results displayed in table 6.6 indicate that the MT-based method utilising SSC is 2.42% better in accuracy than GoEmo. Table 6.7 shows that the difference in accuracy is significant but with a small effect size (<0.10). However, it is interesting to see that the F-score with GoEmo is at 0.648, while SSC has reached only an F-score of 0.541. A closer look at the per-class metrics (table 6.6) reveals a superiority with the GoEmo system at precision (+18.3%), recall (+9.1%) and subsequently at F-score (+10.7%), as compared to the system using SSC. In this context, Abdul-Mageed [3] argues that precision is a more valuable metric when there is a large amount of data, which is the case with the Twitter stream, and the need is to predict users' sentiment with a high precision rather than accurately detecting the sentiment in every coming tweet.

Conclusion: In sum, the MT-based SA system using SSC (MT+SSC) is still significantly better with respect to accuracy than the MT-based system utilising GoEmo (MT+GoEmo) (table 6.7). Thus, the hypothesis that the deep learning architecture of SSC makes it more sensitive to issues like changes in word ordering and loss of contextual information resulting from translation, and hence less effective, which was stated in error analysis (section 6.2.2.3), is actually not confirmed. Nevertheless, the results indicate that the MT+GoEmo system is potentially useful for applications wherein more emphasis is placed on the SA system's precision (correctness within captured/classified tweets).

Future extension: A possible future extension of this investigation is to compare with the projection-based method proposed by Balahur and Turchi [28] (section 6.1). In particular, to explore the scenario of auto-translating an existing publicly available SA corpus of English tweets (e.g. SemEval's data) into Arabic and use

it to train an ML classifier (e.g. SVM). It would be interesting to find out how well this scenario will perform as opposed to the method we have explored in this chapter. However, obtaining gold-standard English translation to benchmark our test-set against will be costly. In addition, we are not aware of an existing Arabic-English Twitter data-set that is manually translated and annotated for SA to be tested against, as in [28].

Metrics	**MT + SSC**		**MT + GoEmo**	
	Pos.	Neg.	Pos.	Neg.
precision	0.268	0.812	0.477	0.847
avg. precision	0.540		**0.723**	
recall	0.347	0.748	0.800	0.558
avg. recall	0.548		**0.639**	
F-score	0.303	0.779	0.598	0.673
avg. F-score	0.541		**0.648**	
accuracy	**66.42**		64.0	

Table 6.6: Comparing MT-based method using SSC vs. GoEmo on Bing translation: results for positive vs. negative

	Positive vs. Negative	
	χ^2 ***(p-value)***	***Effect size***
MT+SSC vs. MT+GoEmo	13.96 (0.000)	0.080

Table 6.7: Comparison between MT-based SA system using: SSC vs. GoEmo on Bing translation with respect to accuracy (stem n-grams) of the independent test-set.

6.3 Summary

The work presented in this chapter is among the first attempts to investigate and empirically evaluate the performance of Machine Translation (MT)-based SA for Arabic tweets. The purpose is to assess how well this tool-based method will perform as compared to data-based methods we have investigated in previous chapters. In particular, we make use of off-the-shelf MT tools, such as Google and Bing translators, to translate Arabic tweets into English. We then use the Stanford Sentiment Classifier (SSC) by Socher et al. [160] to automatically assign sentiment labels (positive, negative or neutral) to translated tweets.

We find that, for subjectivity classification (polar vs. neutral), MT-based SA method attains a reasonable performance of up to 63.10% accuracy that is in line with results reported in previous studies utilising MT-based method on Twitter data (section 6.1). However, this score is not able to compete with our best previous results on this task attained by a lexicon-presence-based DS method at 95% accuracy and by a SL method at 73.99% on word n-grams. An error analysis we conducted (page 175) reveals that an important source of error is the information loss resulting from not translating words that are unknown for the MT tools (e.g. misspelled or dialectal words). Failing to capture informative clues (e.g. misspelled or dialectal) sentiment-bearing features can result in confusing the classifier with a polar instance that seemingly appears as neutral (section 6.2.2.3). Furthermore, SSC was trained on a data-set in which neutral is the minority class (representing around 19%), which might make it less effective in discriminating polar vs. neutral instances.

For sentiment classification, MT-based approaches reach a comparable performance to that attained by more resource intense SA approaches, i.e. hashtag-based DS. As such, MT-based methods have the potential of providing another cheap and effective alternative to building a fully fledged SA system when dealing with under-resourced languages. More specifically, MT-based methods seem to be beneficial for applications with more interest in achieving a high precision rate for identifying positive and negative instances. Although the results of MT-based method are below those achieved with our best performing system on this task that is trained

on a manually labelled gold-standard data-set, the difference in performance can be considered a trade-off against the time and cost otherwise required for obtaining gold-standard labels.

What next? The next chapter summarises the findings of our empirical investigations conducted in this thesis, trying to highlight the main differences among the approaches we investigated.

Chapter 7

Summary of SA Approaches

This chapter first summarises the results of the empirical investigations of chapters 4-6 and critically discusses the main findings. In the second part of this chapter, we present implementation for an SA system for Arabic tweets that exploits our best trained models to automatically label tweets retrieved from the live Twitter stream.

7.1 Results

Our investigations include two different SA approaches: a data-based (including SL and DS) and tool-based (i.e. using existing MT and SA systems). Under the data-based approach we have investigated manual (high quality) vs. automatic (large quantity) methods for obtaining annotated training data for ML classifiers. Table 7.1 summarises results for binary subjectivity (polar vs. neutral) and sentiment (positive vs. negative) and three-way (positive vs. negative vs. neutral) classification for each investigated approach. The table displays the best attained accuracy and F-score and associated feature-set for each set of experiments. Tables 7.2, 7.3 and 7.4 provide ranked lists (with respect to accuracy) of different SA approaches for each sentiment classification task. In the following, we discuss our main findings.

Performance of different feature-sets. We use stem n-grams as the baseline and add individual blocks of features. This has resulted in variable performances across classification tasks and approaches (feature-sets are summarised on page 63).

The word-based n-grams have consistently set a strong baseline, confirming findings of previous SA work [129, 12, 7, 6]. Among the most successful features are morphological, semantic, affective-cues and Twitter-specific feature-sets.

The results indicate that, despite the noise introduced by using MADAMIRA, as a tool designed for MSA only (page 62), the morphological features are still useful for SA on Arabic tweets. This shows the utility of exploiting this feature-set to account for the morphologically-rich nature of Arabic (page 62), as morphological features are amongst the best features across all of the three classification tasks (table 7.1). Previous work on SA in Arabic has either used a small set of POS tags [120], manually extracted a limited set of morphological features [72] or only use a POS feature [8]. Unlike previous work, we employ a rich set of ten auto-extracted morphological features (POS, gender, state, voice, among others), see table 3.13 on page 64, using the publicly available version of the-state-of-the-art Arabic morphological analyser MADAMIRA [131].

For semantic features, we utilise two existing sentiment lexica (MPQA and Arab-Senti, which cover MSA only) and one of our own (DA instances).[1] The results prove this feature-set to be useful not only for SA in MSA [7], but also for SA in social media (i.e. a mixture of MSA and DAs). While the three sentiment lexica we used are manually compiled to account for the aspect of lexicon quality, as strongly stressed by Taboada et al. [165], future investigations might employ a means for careful auto-expansion of lexicon. For this, we have recently presented a system for automatically determining the sentiment orientation of a given instance (single- or multi-word) extracted from Arabic tweets (SemEval'16 Task 7) [143]. Mohammad et al. [119] and Zhu et al. [187] have shown that creating sentiment lexicon comprising Twitter-specific entries both manually and semi-automatically are useful for SA in English tweets.

Another language-dependent (i.e. requires annotated dictionaries) feature-set that found informative is affective-cues (page 65). Affective-cues utilises six binary features that account for the presence of different social signals that can correlate with sentiments (e.g. has-consent, has-laughter and has-prayer). They have shown

[1] Available at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

useful particularly for sentiment classification (positive vs. negative) on the hashtag-based data-set (table 7.1). Although the dictionaries used are relatively small since they are manually created, the usefulness of this feature-set shown by the results suggest the potential of future automatic expansion of the current dictionaries.

The language-style features do not directly contribute to the best performing models (as listed in table 7.1). Nevertheless, they have been shown a considerable degree of success during investigations. For instance, the use of language-style feature-set resulted in a significant gain with the CV setting (page 109). However, the addition of this feature-set resulted in hurting the performance in most of the cases, especially with the independent test-set setting. This is surprising because we anticipated that such features would be helpful to capture patterns (e.g. presence of lengthening and ungrammatical use of punctuations) that can correlate with sentiment and, hence, be informative for the classifiers. However, it seems that the evolving nature of the Twitter stream can result in reducing the utility of the language-style features in this domain, making the detection of (consistent) stylistic patterns not trivial.

Twitter-specific feature-set has shown beneficial for both subjectivity and sentiment analysis (table 7.1). As a language-independent (non-word-based) feature-set, it is characterised by being not sensitive/influenced by issues related to text genres. Instead, it takes advantage from meta-data that is made readily available by Twitter, featuring aspects like whether a tweet is favoured or re-tweeted. Unlike previous work that reported Twitter-specific features to be 'not discriminative' on a small data-set of <2k Arabic tweets [120], our results on 415k auto-labelled tweets (table 7.1) indicate the utility of this feature-set for SA in Arabic tweets. That is, we believe that, with Twitter-specific features, a larger data-set is required in order to infer a pattern/correlation that a classifier can utilise. For instance, our data-sets reveal that: the number of tweets with *is-Retweet:true* tends to be more frequent with positive tweets and *has-hashtag:true* tends to appear more frequently with negative tweets.

Finally, it is interesting to note that the combination of all feature-sets does

not appear among the best recorded scores (table 7.1). A possible explanation is that the presence of some features (e.g. language-style) can hurt the performance. Nevertheless, the combination of all blocks of features resulted in performance gain during investigations. For instance, the combination of all blocks of features has resulted in a significant gain of 11.09% accuracy over the stem n-grams baseline with the GS1 data-set (page 91). As such, we conclude that utilising individual blocks seem to be more successful for SA in Arabic than experimenting with all feature-sets combined. Wilson et al. [174], in contrast, found that the combination of all features (e.g. syntactics, semantic, POS, among others) is the best for SA in English. Previous work on Arabic has either investigated word-based (syntactic) features only [13, 9, 60, 122], reported the best attained scores on a smaller and/or selected-dialect data [120, 64] or used subset of the features we utilised [8]. Our results also show that the choice of which feature-set depends on the approach utilised (SL or DS) and on the classification task (subjectivity or sentiment classification; binary or three-way).

Data quantity vs. quality. An important aspect for ML classifiers is the trade-off between data quantity (noisy auto-labelling) vs. quality (manual-labelling). Our work involves investigating SL (manual-labelling-based) and DS (auto-labelling-based) approaches. The results indicate each one of these methods (i.e. SL or DS) can be more suitable for the performance of one classification task (i.e. sentiment or subjectivity) than the other. To illustrate, for subjectivity classification (polar vs. neutral), the results show the utility of exploiting large and automatically labelled data, with the lexicon-presence-based DS method in the lead on this task, and emoticon-based DS is the second best method (table 7.2). This indicates the usefulness of adapting a DS approach as opposed to manually annotating a corpus for this task. A possible explanation is that discriminating polar vs. neutral instances is expected to be an easier task, because the polar class is predominantly in DAs while the neutral class is predominantly in MSA (see figure 5.1 on page 135). This allows the subjectivity classifiers to infer different linguistic patterns and lexical variation to distinguish between the two classes.

As for sentiment analysis, the results show that the quality of the sentiment

		Polar vs. Neutral		**Positive vs. Negative**		**Positive vs. Negative vs. Neutral**	
	Approach	F	Acc.	F	Acc.	F	Acc.
Data-based	**SL** (GS1+GS2) *(feat-set)*	0.735 *(stem)*	73.99* *(stem)*	**0.780** *(twt-specific)*	**77.97** *(morph)*	0.641 *(semant.)*	64.10* *(semant.)*
	DS (emoticon-based) *(feat-set)*	0.950 *(morph)*	95.19 *(morph)*	0.590 *(morph)*	64.82* *(morph)*	0.704 *(stem)*	69.67* *(stem)*
	DS (hashtag-based) *(feat-set)*	0.581 *(twt-specific)*	62.58* *(twt-specific)*	0.674 *(Affec.-cues)*	69.58* *(Affec.-cues)*	0.413 *(morph)*	43.81* *(morph)*
	DS (lexicon-pres.-based) *(feat-set)*	**0.953** *(stem)*	**95.36** *(stem)*	0.574 *(twt-specific)*	56.21* *(twt-specific)*	**0.710** *(stem)*	**71.57** *(stem)*
	DS (lexicon-aggreg.) *(feat-set)*	0.910 *(stem)*	91.11* *(stem)*	0.543 *(twt-specific)*	53.60* *(twt-specific)*	0.630 *(stem)*	64.96* *(stem)*
Tool-based	**MT-based** (Google + SSC)	0.505	56.64*	0.509	60.01*	0.420	46.30*
	MT-based (Bing + SSC)	0.558	63.10*	0.553	66.42*	0.450	50.32*

Table 7.1: Benchmarking different SA systems on the independent test-set. * denotes a statistically-significant difference vs. **the best score** ($p<0.05$).

annotation is more important than quantity for discriminating positive vs. negative instances [143]. Table 7.3 shows SL to attain the best performance on this task, significantly outperforming much larger, auto-labelled DS data-sets (emoticon- and lexicon-based). Error analysis and manual examinations of samples of misclassified tweets revealed multiple sources of difficulties with automatic sentiment annotation of positive and negative instances. For instance, the emoticon-based DS investigations have shown that detecting the positive class is more challenging due to the misleading use of emoticons, i.e. mistyped or sarcastic (page 138). Unlike the emoticon- and lexicon-based DS methods, the results of the hashtag-based DS method rank second on this task (table 7.3). This suggests that the hashtag-based DS method can provide training data with a better quality compared to emoticon- and lexicon-based DS for positive vs. negative classification. Although the SL data is 8% significantly better than the hashtag-based DS data on this task, the difference in performance can be considered a trade-off for the cost that would otherwise be required to obtain gold-standard labels.

In sum, with limited or less-resourced languages, investing in creating high quality linguistic resources (e.g. manually-annotated SA corpora) is likely to help training SA classifiers with promising performance that is comparable to the state-of-the-art SA systems in a well-resourced language, e.g. English (page 116). However, there is a possibility that such resources can become less effective over time, especially with Twitter data [61]. To this end, auto-labelling methods can provide a cheap and fast alternative for obtaining training data, but, as a trade-off, the sentiment labels are noisy and this can come at the cost of training a less effective classifier. In section 7.2, we describe our attempt to combine the two approaches (auto- and manual-labelling) to create a system for automatically predicting sentiments of tweets retrieved from the live Twitter stream.

Data-based vs. tool-based approaches. Another direction in our investigations is the use of a tool-based method in which we explored the scenario of what if there is no annotated data readily available? An existing approach to address this question is by leveraging resources from a well-studied language like English (page

	Polar vs. Neutral					
	Data size	Prec.	Recall	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Lexicon-presence DS	78.5k	0.954	0.951	94.817 (0.000)	0.307	0.0480
Emoticon-based DS	121.6k	0.953	0.949	98.572 (0.000)	0.315	0.0511
Lexicon-Aggreg. DS	83.2k	0.843	0.818	137.329 (0.000)	0.370	0.0891
SL (GS1+GS2)	8k	0.774	0.780	530.403 (0.000)	0.387	0.2603
MT-based (Bing+SSC)	215.2k	0.557	0.540	662.08 (0.000)	0.432	0.6391

Table 7.2: A ranked list (accuracy) for SA approaches on polar vs. neutral task.

162). Unlike data-based methods in which annotated data is obtained (manually or automatically) to train ML classifiers and build an SA system from scratch, we used a tool-based method in which unannotated Arabic instances are automatically translated to English (e.g. using Bing) and annotated using off-the-shelf SA systems for English. Because of the present challenges with publicly available MT tools (e.g. mistranslation resulting from DAs and misspellings, see page 175), the tool-based system performs significantly worse than data-based systems both for subjectivity and sentiment analysis (table 7.1). Nevertheless, we observed that the tool-based data can attain a comparable performance to that achieved by the best data-based method for positive vs. negative with respect to precision, i.e. correctness within classified instances (table 7.3). This is interesting, considering the cost of annotating new data and building a system from scratch (i.e. as it is the case with data-based methods).

In sum, with public resources currently available for Arabic, it seems that a tool-based method is more suitable for sentiment classification than subjectivity classification, and it is more appropriate for systems with more emphasis on precision rate rather than accuracy or recall rates.

	Positive vs. Negative					
	Data size	Prec.	Recall	χ^2 ***(p-value)***	***Effect size***	***Classification error***
SL (GS1+GS2)	3.7k	0.789	0.776	37.480 (0.000)	0.132	0.2377
Hashtag-based DS	130.2k	0.690	0.613	118.466 (0.000)	0.232	0.3078
MT-based (Bing+GoEmo)	1.6M	0.723	0.639	153.28 (0.000)	0.264	0.3357
Emoticon-based DS	1.2M	0.498	0.497	54.644 (0.000)	0.157	0.4722
Lexicon-presence DS	415.8k	0.584	0.589	740.41 (0.000)	0.582	0.4571
Lexicon-Aggreg. DS	487.5k	0.584	0.582	1159.97 (0.000)	0.728	0.4860

Table 7.3: A ranked list (accuracy) for SA approaches on positive vs. negative task.

	Positive vs. Negative vs. Neutral					
	Data size	Prec.	Recall	χ^2 ***(p-value)***	***Effect size***	***Classification error***
Lexicon-presence DS	78.5k	0.746	0.701	104.883 (0.000)	0.316	0.2842
Emoticon-based DS	121.5k	0.721	0.684	102.138 (0.000)	0.325	0.3032
Lexicon-Aggreg. DS	83.2k	0.624	0.640	176.261 (0.000)	0.427	0.3503
SL (GS1+GS2)	8k	0.651	0.641	123.926 (0.000)	0.187	0.3660
MT-based (Bing+SSC)	215.2k	0.420	0.419	1956.6 (0.000)	0.743	0.4968

Table 7.4: A ranked list (accuracy) for SA approaches on positive vs. negative vs. neutral.

7.2 A System for Sentiment Analysis of Arabic Tweets (SAAT)

The last element in the framework presented in section 2.5 (page 36) is to develop a system that deploys the best performing trained models. This section describes a system for Sentiment Analysis of Arabic Tweets (SAAT) that retrieves tweets from the live Twitter stream about given queries and utilises our best trained models to automatically assign retrieved tweets with sentiment labels.

SAAT follows a hierarchical structure (figure 7.1), i.e. two-level binary classification, because our investigations revealed that hierarchical systems have yielded better results than a flat or single-level three-way classification (page 93). SAAT utilises our best trained models: lexicon-presence-based DS model for subjectivity classification and fully-supervised-learning SL (GS1+GS2) model for sentiment classification (figure 7.1).

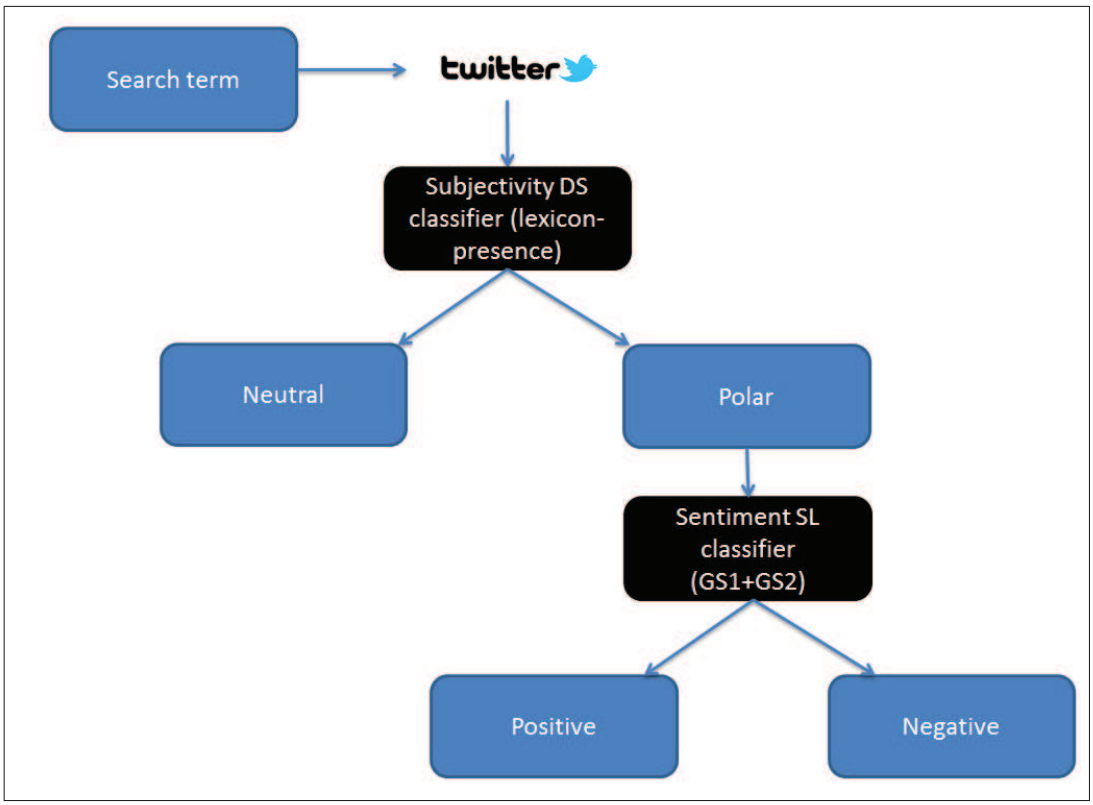


Figure 7.1: System architecture of SAAT.

SAAT follows the mechanism adopted in the Sentiment140 system by Go et al.

[81],[2] as one of the top performing publicly available SA systems for English tweets [2]. This mechanism involves querying the live Twitter stream and automatically annotating retrieved tweets as positive, negative or neutral, using pre-trained ML models. SAAT works as follow:

1. Executing SAAT.jar file will prompt a text message requesting a query word (see figure 7.2).

```
bash-4.1$ java -jar "SAAT.jar"
Enter a query word: بمارت
بمارت   1
tweet number 5
waiting...
tweet number 14
waiting...
tweet number 21
waiting...
tweet number 29
waiting...
tweet number 33
waiting...
tweet number 45
waiting...
tweet number 52
waiting...
tweet number 55
waiting...
```

Figure 7.2: SAAT snapshot1: Sending a query via SAAT to search the live Twitter stream for tweets about *'Trump'*.

2. Each retrieved tweet from the live Twitter stream is saved as: 1) tweet text (cleaned up), following the steps described in section 3.2 (page 56); and 2) tweet's JSON object with all properties of this tweet (page 41).

3. Tweets saved into the first file (cleaned up) will then be passed into a subjectivity model. For this, we use the classifier that was trained on the lexicon-presence-based DS data-set due to its superior performance on this task. The output here is a file with the retrieved tweets classified as polar or neutral (see figure 7.3).

[2] `http://www.sentiment140.com/`

4. Polar tweets will then be passed to a sentiment classification model. For this, we use the classifier that was originally trained on SL (GS1+GS2) data. The output here is the subjective tweets being classified as positive or negative (see examples in table 7.5).

5. As additional options, the current source code of SAAT allows for: 1) setting a set of queries in a separate text file to search Twitter for each entity/query in sequence; 2) setting up a maximum number of tweets that can be retrieved for a single query; 3) sending Arabic or English queries. The latter will result in collecting Arabic tweets containing the specified English query word (see page 212).

The source code, trained models and data-sets are freely-available. [3]

To assess how well SAAT will perform on live stream data, we used SAAT to retrieve and automatically annotate tweets about various topics. After duplicates removal, the resulting data-set includes a total of 34,829 Arabic tweets. The tweets were randomised, auto-labels were removed, and a random sample of 500 tweets were assigned to one of our human annotators to classify them as positive, negative or neutral. Again, tweets with *mixed* emotions were classified based on the strongest emotion conveyed and tweets with unclear sentiment orientation were labelled as *uncertain* (see page 43). Following this, tweets identified as *uncertain* were excluded, resulting in a total of 405 tweets. The contingency table 7.6 displays the results of human annotation and automatic annotation. Overall accuracy is at 65.68% (see table 7.7). Although this is lower than the performance of our best performing system with a hierarchical structure at 75.11% on our independent test-set (page 112), this result reflects a promising performance considering the time lag between data used to train models and these 405 auto-labelled tweets. Furthermore, this is in line with the best performing accuracies on English tweets ranging between 65-71% [2]. F-scores in SemEval-2015 range between 0.648 and 0.248 for SA on English tweets [145]. Overall, the performance drop of our best models when used on live tweets confirms the need for continuously obtaining training data (e.g. using

[3]Available at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

hashtags) in order to keep models up-to-date. In next chapter, we discuss the future possibilities for utilising incremental learning to address this issue.

```
retrieved tweets successfully written to output file...



Annotating retrieved tweets as polar vs. neutral (get predictions/labels)
***********************************************
label what? بمارت tweets

load pre-trained serialized model ...

Experiment date/time: 2016/03/15 15:00:09


load unlabeled data ...
load and deserialize pre-trained model (polar vs. neutral) ...

label the new instances ...

labeled data about بمارت successfully written into an ARFF ...




Annotating subjective tweets as positive vs. negative  (get predictions/labels)
***********************************************

load pre-trained serialized model ...



load subjective data ...
load and deserialize pre-trained model (positive vs. negative) ...

label the subjective instances ...

labeled data about بمارت successfully written into an ARFF ...

Total positive : 2

Total negative : 8

Total neutral : 32

Total annotation time: 8 Seconds


Working Directory = /home/eaar1
-bash-4.1$
```

Figure 7.3: SAAT snapshot2: The retrieved tweets about *'Trump'* are saved into an output file before being classified as polar or neutral. Next, the polar tweets are classified as positive or negative.

Potential of SAAT: Being able to retrieve tweets from the live Twitter stream and automatically assign them with sentiment labels, SAAT is potentially useful to serve a wide range of real-world applications for SA (page 2). For instance,

1	لو سَاندرز عمل معجزه وقَابل ترَامب، فَاَالَامر محسوم لهيلَاري ***If Sanders gets to the final with Trump, then things will be working really well for Hillary.***	positive
2	انونيموس تعلن حربَا شَامله علي دونَالد ترَامب ***Anonymous hackers declare total war on Donald Trump.***	negative
3	هل دونَالد ترَامب جورج وَاليس الجديد؟ ***Is Donald Trump the George Wallace of this time?***	neutral

Table 7.5: Examples of tweets about *'Trump'* auto-labelled via SAAT.

		Auto-Annotation			
		Negative	Positive	Neutral	Total
Manual annotation	Negative	87	8	78	173
	Positive	10	71	32	113
	Neutral	10	1	108	119
	Total	107	80	218	405

Table 7.6: Contingency table of a random sample of 405 tweets along with their auto-annotation via SAAT and manual annotation.

Metrics	**Pos.**	**Neg.**	**Neut.**
precision	0.887	0.813	0.495
avg. precision	0.732		
recall	0.283	0.451	0.907
avg. recall	0.547		
F-score	0.429	0.580	0.641
avg. F-score	0.550		
accuracy	65.68		

Table 7.7: SAAT results on a random sample of 405 tweets.

politicians can use such system to monitor how the general public currently feel towards them (e.g. after a public statement or new policy launch).

Following Sentiment140 [81], the current first version of SAAT uses only word-based features to classify tweets. The purpose is to create a system that simulates Sentiment140 (which currently covers English and Spanish) but for Arabic. Future expansions of the system will involve extracting more feature-sets (e.g. embedding MADAMIRA to extract morphological features for the retrieved tweets).

A version of this system has won SemEval-2016 Task 7 (Arabic Twitter subtask) [143, 106]. Here, modifications were made to accommodate the task description: predicting sentiment intensity scores (i.e. 0-1) instead of sentiment labels (e.g. positive or negative).

7.3 Summary

This chapter presents a summary of the empirical investigations conducted throughout this work. The results reveal that the choice between data quality (manual-annotation-based approaches) and data quality (automatic-annotation-based approaches) is a task dependent. That is, the results show that data quantity (e.g. using a lexicon-based DS) approach to automatically obtain noisy labels is more useful for subjectivity classification (polar vs. neutral). A possible explanation is that polar tweets tend to be dialectal while neutral tweets tend to be in MSA. In sentiment classification (positive vs. negative), where positive and negative tweets can have a similar degree of dialectness, the results indicate that better data quality is more useful. For instance, we found that the use of emoticons to automatically annotate Arabic tweets as positive or negative can be misleading, with many emoticons being mistyped or used sarcastically.

We also looked into employing a data-based vs. tool-based approach for SA on Arabic tweets. For the latter, we evaluate the performance of Machine Translation (MT)-based that make use of off-the-shelf MT tools (e.g. Bing) and SA systems for English. We found that tool-based approaches can provide a cheap and effective alternative to building a system from scratch but seem to be more effective for sentiment than subjectivity classification. The choice between data-based and tool-based might depend on the intended application. For instance, if the intended application requires placing more emphasis on precision rate and with no annotated data readily available, then tool-based is expected to be beneficial and cheaper.

This chapter also describes the first steps towards creating an SA system for Arabic tweets. The system follows a hierarchical structure and utilises our best performing model at each level. The system is capable of retrieving tweets from the Twitter stream about a certain query and classifying them as positive, negative or neutral. SAAT is anticipated to be a valuable tool, especially that very few SA systems have been made available to the public [164], none of them is for Arabic. A version of this system has won SemEval-2016 Task 7 (Arabic Twitter subtask), out of 3 competing systems.

Chapter 8

Conclusion and Future Work

Sentiment Analysis (SA) is a text classification task that concerns the automatic extraction and identification of sentiment-related information from a given text instance into the sentiments they convey, i.e. positive, negative or neutral [167]. There is a growing interest in recent years in studying sentiments conveyed in user-generated text, which is coincided with the increasing prevalence of social media. Popular social media platforms, such as Twitter, are used by an extremely large number of users as a means of communication through short messages that convey personal opinions, attitudes, emotions, preferences, and so on. SA provides means for automatically summarising sentiments expressed in text. Common applications of SA include assessing a product/service's success, anticipating financial performance in the stock market and as a tool used by political analysts (e.g. for detecting popularity of political candidates/parties).

Research gap: In this work, we have investigated and identified shortcomings in SA for Arabic. Compared to English, research on SA for Arabic social media is still limited. A main reason for this is the limited availability of linguistic resources (i.e. annotated data-sets and subjectivity lexica) for SA. Despite recent interesting efforts to build web corpora for SA in Arabic (page 19), none of them had been made publicly available by the time of this thesis. In addition, existing SA work for Arabic has focused on domains like: reviews, newswire and web forums [18, 7, 1]. Less work has studied SA in the noisy genre of social media. Previous work

on SA of Arabic tweets suffers from: small data-sets (up to 3k tweets); narrow feature-sets employed; evaluating data without consideration of the dynamic nature of Twitter. More importantly, previous work was carried out without comparing different existing approaches, techniques and feature-sets against a benchmark test-set. This is required to gain a better understanding of how these factors influence performance of an SA system.

Thesis goals: The main goals of this work are:

1. to empirically investigate and evaluate current SA techniques for Arabic (as an under-resourced language) and identify issues specific to the Arabic language;
2. to determine the influence of feature-sets, data quantity vs. quality on the models' performance;
3. the provision and use of freely available data and tools.

8.1 Main Conclusions of Empirical Investigations

The empirical investigations presented in this work use three main approaches and utilise a variety of feature-sets to automatically determine sentiments conveyed in Arabic tweets. We benchmark various existing approaches for SA on an independent and diverse test-set of >3.5k instances, collected at different points in time, following SemEval [146]. We assess the effects of feature-sets and data quality vs. quantity on SA performance.

Supervised Learning (SL) Approaches: First, we explore a fully-supervised machine learning approach that uses a gold-standard manually-annotated data-set (chapter 4). We demonstrate the superiority of our extended feature-sets, outperforming previous work with an accuracy improvement of 2.65% on subjectivity classification and 9.42% on sentiment analysis using the data-set used by Mourad and Darwish [120]. The most beneficial feature-set is the morphological feature-set that is automatically-extracted using the publicly available version of the state-of-the-art morphological analyser, showing the utility for employing features accounts

for the morphology-rich nature of Arabic. Other successful feature-sets include semantic (e.g. presence of positive/negative lexicon) and affective cues features (e.g. presence of laughter, sigh, prayers and consent) that utilise manually created dictionaries. Twitter-specific, as a language-independent features also found among the most informative feature-sets (e.g. is-retweeted). Amongst the least informative features come the language-style features (e.g. presence of lengthening and ungrammatical use of punctuations). A possible reason is the difficulty of detecting a consistent correlation between a stylistic pattern and sentiments due to the evolving nature of Twitter [61]. Our SL classifiers attained a top $F_{(positive,negative)}$ score at 0.612, which is comparable to the best results reported on English tweets in SemEval'15 with an $F_{(positive,negative)}$ at 0.648 [145].

Furthermore, experiments in chapter 4 revealed a notable impact of topic shift issues associated with the Twitter stream data on the performance of classifiers. That is, using a standard cross-validation evaluation setting, the fully-supervised SA systems were able to attain an average accuracy score of 87.89%. However, re-evaluating SA systems against our independent test-set resulted in a performance drop of 24.13% accuracy. We investigated the hypothesis that models do not transfer well because of the topic shifts issue, especially on Twitter data, and the prominent role of word n-grams features in our models. Thus, we utilised larger training data with which the performance gap has been reduced from 24.13% to 4.93% using a data-set 4 times larger. We concluded that more data can improve performance on the independent test-set. However, continuously obtaining gold-standard sentiment labels is costly. Therefore, we turned to systematically investigate and evaluate solutions previously used in literature. Specifically, we studied the utility of exploiting readily available features (e.g. emoticons) as 'noisy' labels, using *Distant Supervision* (DS) approaches.

Distant Supervision (DS) Approaches: Investigations presented in chapter 5 covered two DS approaches: Twitter's conventional-markers-based DS and lexicon-based DS approaches. Our results suggest data quality (manual-labelling) vs. quantity (automatic-labelling) aspect is task dependent. We found that DS approaches

perform well on subjectivity analysis (polar vs. neutral) with an accuracy score of up to 95%. However, we anticipated that this highly optimistic performance can be partially attributed to the fact that the neutral class in training and test data is predominantly in MSA, while the polar class is mostly in DAs. Therefore, it seems that the models mostly learnt to distinguish MSA vs. DAs. Subsequently, we assessed the performance of the best performing DS models on tweets retrieved from the live Twitter stream (results are displayed in table 7.7 on page 198). The classifier was able to achieve an accuracy score of 73.01% for subjectivity classification, which is still a reasonable performance and comparable to the scores reported in previous work on this task for English at 75.3% accuracy [174] and outperforming previous work on Arabic tweets at 63.6% [120] and 71.38% [6].

As for sentiment analysis (positive vs. negative), the DS methods (with a >6 times larger training-set than SL ones) was not able to outperform the best score attained by SL models on the independent test-set for this task at 77.97% accuracy. A second round of investigations was conducted in order to boost the performance of DS methods on binary SA (positive vs. negative). This involved collecting up to 9 times larger emoticon-based data than previous emoticon-based data and exploiting hashtags to collect a new training data. The results show that using a larger emoticon-based DS data has resulted in an accuracy improvement of 2.56% as compared to previous (smaller) emoticon-based data-set. An interesting finding is that the hashtag-based DS data (130.2k) attain an accuracy score of up to 69.58%, which is 4.76% better than the best accuracy achieved using the extended emoticon-based data-set (1.2M). We conclude that hashtags are more reliable for labelling sentiments automatically, as they seem to introduce less noise.

With respect to the lexicon-based DS data-sets, the best accuracy performance for positive vs. negative is attained at 56.21%, which is significantly lower than the score attained by the best SL classifier at 77.97% and by the hashtag-based DS classifier at 69.58%. Interestingly, experimenting with a 5 times larger lexicon-based training data-set has not yielded any improvements. Instead, it slightly hurts performance of SA classifiers. As such, we concluded that adding more data is

beneficial for SA classifiers, but not with all approaches, i.e. only if labels are not too noisy. Overall, we found that DS introduces different levels of noise.

Machine Translation-based (MT) Approaches: Subsequently, we explored in chapter 6 the use of a Machine-Translation (MT)-based method for SA that exploits existing tools for English. The MT-based approach uses a publicly accessible MT tool (e.g. Google or Bing) to translate Arabic tweets to English and employs an off-the-shelf SA system for English (e.g. the Stanford Sentiment Classifier). The results indicate MT-based method as a viable, fast and cheap alternative to building SA systems from scratch, when no annotated data of sufficient quality and quantity is readily available. The best recorded accuracy scores with this approach are at 63.10% for polar vs. neutral and at 66.42% for positive vs. negative. Both scores are significantly lower than our best results for subjectivity classification (with lexicon-based DS) and sentiment analysis (with gold-standard SL) methods. However, we observed that the precision rate for sentiment classification is at 0.723 for positive vs. negative, as compared to the best score attained by gold-standard SL at 0.789. This suggests the utility of the MT-based SA approach we explored, especially for applications that put more emphasis on precision rather than accuracy or recall. An error analysis revealed mistranslation owing to misspelled or dialectal words as a main source of error in this approach.

A System for Sentiment Analysis in Arabic (SAAT): Finally, we presented an SA system for Arabic, namely SAAT. SAAT utilises our best performing models. The system follows a hierarchical two-level binary classification structure, as our results shows a superiority for this design over the flat three-way classification. SAAT retrieves tweets from the Twitter stream about a given query and classifies them as polar or neutral. Polar tweets are then classified as positive or negative. Using SAAT, we collected and automatically annotated a set of >34k tweets.[1] We manually annotated a random sample of 405 tweets and recorded an accuracy score of 65.68% across positive, negative and neutral.

[1]The auto-labelled data-set is freely available at: `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`

8.2 Contributions

Automatically determining the sentiment contained in highly noisy text, such as tweets, is an important text classification problem that has become an active research area mainly due to its numerous real-world applications. This thesis focuses on SA in Arabic, as a less-resourced and morphologically-complex language, and contributes the following:

1. Systematically evaluating and comparing existing approaches to SA for Arabic:
    - We find that the following feature-sets lead to a significant performance boost:
        - Morphological features.
        - Semantic features.
        - Affective-cues/social-signals features.
        - Twitter-specific features.
    - We find that increasing by 4 times the size of the training data-set for SL leads to a significant improvement of 10.37% and a significant reduction in performance gap on the independent test-set from 24.13% to only 4.93%.
    - We find that there is a trade-off for DS approaches between data quality and quantity, with the conventional marker (hashtag) approach being the least noisy.
    - We find that using a combination of MT and existing publicly available SA systems for English can eliminate the need for data annotation for sentiment classification, with a promising precision rate of 0.723.
2. We release publicly available data-sets:[2]
    - **Data-set1:** 9k of gold-standard (manually annotated) tweets, which has been released via an ELRA repository.[3] The corpus by far has been accessed 162 times and downloaded more than 110 times. It has been

[2] `http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html`
[3] Available at: `http://www.resourcebook.eu/shareyourlr/index.php#`

also used by other research, e.g. Talaat et al. [104], Salameh et al. [153] and Htait et al. [96].

- **Data-set2:** 1.2M tweets automatically labelled for sentiments using positive/negative emoticons.
- **Data-set3:** 130.2k tweets automatically labelled for sentiment using sentiment-bearing hashtags.
- **Data-set4:** 3.5k benchmark test-set of gold-standard manually labelled Arabic tweets.
- **Data-set5:** 34k tweets automatically labelled using our SA system SAAT.

We also publicly shared the following sentiment lexica:

- **Sentiment-lexicon1:** A manually annotated dialectal subjectivity lexicon of 489 items, which has been used and automatically expanded by Salameh et al. [153].[4]
- **Sentiment-lexicon2:** An automatically translated and manually filtered MPQA lexicon [173] of 2,852 items.

3. We release a publicly available SA tool for Arabic tweets, which combines the best trained models. A version of this system has won the SemEval-2016 Task 7, Arabic Twitter subtask [143, 106].

8.3 Future Directions

Possible directions of future work may include:

- **Experimenting with topic-relevant SA and systematically compare it to approaches presented in this thesis.** The importance of this extension is in filtering tweets for topic relevance that would make determining sentiments towards a specific topic of interest more accurate. This is because Dacres et al. [51] found that a keyword-based method, like the one we used

[4] `http://saifmohammad.com/WebPages/ArabicSA.html`.

in this work, can be be 'too broad' to accurately capture tweets referring to a particular topic.

- **Investigating new and updated releases of Arabic tools** used to build and extract features-sets for learning SA classifiers. NLP on dialectal Arabic is an active research area nowadays with lots of interesting efforts (e.g. to build corpora, morphological analysers, and MT systems), as discussed in detail in chapter 2 (page 19). For instance, Pasha et al. [131] promised further expansions in upcoming releases of MADAMIRA, such as CODAfy [86]. CODAfy is a component that attempts to enforce certain orthographical conventions, i.e. imposing orthography standardisation on dialectal text, which can be a useful pre-processing tool. In addition, a tool like ELISSA that translates/maps dialectal text instances into MSA, once released, can help alleviating noise and data sparsity issues caused by DAs (page 23). We expect an improvement in performance of the morphological and semantic feature-sets due to a reduction in noise.

- **Compare incremental learning as opposed to the batch learning used in this work.** Batch learning is the mode of training wherein an ML model learns once, i.e. resultant models are not updatable. Guerra et al. [82] argue that data stream classification tasks require the evolving nature of the stream to be dealt with, i.e. concept/topic drift issues. This involves employing a means for constantly updating the classification models, e.g. [38, 39].

- **Investigating alternatives for combining multiple classifiers, rather than choosing the best one.** For instance, *bagging* allows an ensemble of classifiers to vote for a sentiment label [43]. Thus, the classification decision is made by an ensemble of classifiers, allowing for performance to be further enhanced by combining the strengths of more than one classifier.

Bibliography

[1] A. Abbasi, H. Chen, and A. salem. Sentiment analysis in muliple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(1-34), 2008.

[2] A. Abbasi, A. Hassan, and M. Dhar. Benchmarking twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[3] M. Abdul-Mageed. *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. PhD thesis, The School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States, 2015.

[4] M. Abdul-Mageed and M. Diab. AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[5] M. Abdul-Mageed and M. Diab. SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2014.

[6] M. Abdul-Mageed, M. Diab, and S. Kübler. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1):20–37, 2014.

[7] M. Abdul-Mageed, M. T. Diab, and M. Korayem. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] M. Abdul-Mageed, S. Kuebler, and M. Diab. SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. Association for Computational Linguistics, 2012.

[9] N. Abdulla, N. Ahmed, M. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference*, pages 1–6. IEEE, 2013.

[10] N. Abdulla, S. Mohammed, M. Al-Ayyoub, M. Al-Kabi, et al. Automatic lexicon construction for arabic sentiment analysis. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pages 547–552. IEEE, 2014.

[11] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, M. Al-Ayyoub, M. N. Al-Kabi, and S. Al-rifai. Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, 9(3):55–71, 2014.

[12] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[13] S. Ahmed, M. Pasquier, and G. Qadah. Key issues in conducting sentiment analysis on Arabic social media text. In *9th International Conference on Innovations in Information Technology (IIT), 2013*, pages 72–77. IEEE, 2013.

[14] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, pages 39–50. Springer, 2004.

[15] M. Al-Badrashiny, M. Diab, N. Habash, M. Pooleery, O. Rambow, and R. Roth. *MADAMIRA v1.0 User Guide*. Center for Computational Learning Systems, Columbia University, 2014.

[16] S. Al-Osaimi and K. M. Badruddin. Role of emotion icons in sentiment classification of Arabic tweets. In *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*, pages 167–171. ACM, 2014.

[17] R. Al-Sabbagh and R. Girju. YADAC: Yet another dialectal Arabic corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[18] M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider. Human annotated Arabic dataset of book reviews for aspect based sentiment analysis. In *3rd International Conference on Future Internet of Things and Cloud (FiCloud), 2015*, pages 726–730. IEEE, 2015.

[19] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman. Subjectivity and sentiment analysis of Arabic: Trends and challenges. In *IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), 2014*, pages 148–155. IEEE, 2014.

[20] L. Albraheem and H. S. Al-Khalifa. Exploring the problems of sentiment analysis in informal Arabic. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, pages 415–418. ACM, 2012.

[21] A. Alhothali and J. Hoey. Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 1548–1558, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[22] B. S. AlMutawa. Automatic emotion and dialect detection tool for Arabic language. Master's thesis, The School of Electronic Engineering and Computer Science, Queen Mary University of London, 2013.

[23] S. Amir, W. Ling, R. Astudillo, B. Martins, M. J. Silva, and I. Trancoso. INESC-ID: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618, Denver, Colorado, June 2015. Association for Computational Linguistics.

[24] Arab-Social-Media-Report. Twitter in the Arab Region. `http:http://www.arabsocialmediareport.com/Twitter`. Accessed: 2015-03-12.

[25] Arab-Social-Media-Report. Civil Movements: The Impact of Facebook and Twitter. `http://unpan1.un.org/intradoc/groups/public/documents/dsg/unpan050860.pdf`, 2011. Accessed: 2015-03-12.

[26] M. A. Attia. *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation.* PhD thesis, School of Languages, Linguistics and Cultures, University of Manchester, UK., 2008.

[27] A. Balahur, R. Mihalcea, and A. Montoyo. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1):1–6, 2014.

[28] A. Balahur and M. Turchi. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 49–55, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA.

[29] J. Baldridge and M. Osborne. Active learning and the total cost of annotation. In *EMNLP*, pages 9–16, 2004.

[30] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135. Association for Computational Linguistics, 2008.

[31] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics, 2001.

[32] F. Barbieri and H. Saggion. Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, 2014.

[33] M. Barthel, E. Shearer, J. Gottfried, and A. Mitchell. The Evolving Role of News on Twitter and Facebook. `http://www.journalism.org/2015/07/14/the-evolving-role-of-news-on-twitter-and-facebook/`, 2015. Accessed: 2016-02-06.

[34] V. Basile and M. Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, 2013.

[35] R. Batuwita and V. Palade. Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications*, pages 83–99, 2013.

[36] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[37] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.

[38] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive online analysis. *The Journal of Machine Learning Research*, 99:1601–1604, 2010.

[39] A. Bifet, G. Holmes, and B. Pfahringer. *MOA-TweetReader: Real-Time Analysis in Twitter Streaming Data*, pages 46–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[40] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[41] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint arXiv:0911.1583*, 2009.

[42] U. M. Braga-Neto. Classification and error estimation for discrete data. *Current genomics*, 10(7):446–462, 2009.

[43] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[44] R. Buettner and K. Buettner. A systematic literature review of twitter research from a socio-political revolution perspective. In *the 49th Hawaii International Conference on System Sciences (HICSS)*, pages 2206–2215. IEEE, 2016.

[45] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[46] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[47] P. Chaovalit and L. Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE, 2005.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[49] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[50] R. Cotterell and C. Callison-Burch. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[51] S. Dacres, H. Haddadi, and M. Purver. Topic and sentiment analysis on OSNs: a case study of advertising strategies on twitter. In *Proceedings of the 6th ASE International Conference on Social Computing*, May 2014.

[52] K. Darwish and W. Gao. Simple effective microblog named entity recognition: Arabic as an example. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[53] K. Denecke. Using SentiWordNet for multilingual sentiment analysis. In *IEEE 24th International Conference on Data Engineering Workshop, 2008. ICDEW 2008.*, pages 507–512. IEEE, 2008.

[54] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

[55] B. S. Dhakar, S. K. Sinha, and K. K. Pandey. A survey of translation quality of English to Hindi online translation systems (Google and Bing). *International Journal of Scientific and Research Publications*, 313, 2013.

[56] M. Diab and N. Habash. Natural language processing of Arabic and its dialects. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP). Doha, Qatar. October 25-29, 2014*, page 10. Association for Computational Linguistics (ACL), 2014.

[57] M. Diab, K. Hacioglu, and D. Jurafsky. Automated methods for processing Arabic text: From tokenization to base phrase chunking. *Arabic*

Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer, 2007.

[58] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2015.

[59] K. Duh, A. Fujino, and M. Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 429–433. Association for Computational Linguistics, 2011.

[60] R. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. Sentiment analysis in Arabic tweets. In *5th International Conference on Information and Communication Systems (ICICS), 2014*, pages 1–6, April 2014.

[61] J. Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369, 2013.

[62] S. R. El-Beltagy and A. Ali. Open issues in the sentiment analysis of Arabic social media: A case study. In *9th International Conference on Innovations in Information Technology (IIT), 2013*, pages 215–220. IEEE, 2013.

[63] A. El-Halees. Arabic opinion mining using combined classification approach. In *the Proceedings of the International Arab Conference on Information Technology (ACIT'2011), December 11-14*, Riyadh, Saudi Arabia, 2011. Naif Arab University for Security Sciences.

[64] N. El-Makky, K. Nagi, A. El-Ebshihy, E. Apady, O. Hafez, S. Mostafa, and S. Ibrahim. Sentiment analysis of colloquial Arabic tweets. *Academy of Science and Engineering, USA*, 2015.

[65] H. El-Sahar and S. R. El-Beltagy. A fully automated approach for Arabic slang lexicon extraction from microblogs. In *Computational Linguistics and Intelligent Text Processing*, pages 79–91. Springer, 2014.

[66] H. Elfardy, M. Al-Badrashiny, and M. Diab. AIDA: Identifying code switching in informal Arabic text. *EMNLP 2014*, page 94, 2014.

[67] H. Elfardy and M. T. Diab. Sentence level dialect identification in Arabic. In *ACL (2)*, pages 456–461, 2013.

[68] R. Eskander and O. Rambow. SLSA: A sentiment lexicon for Standard Arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2550, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[69] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[70] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[71] A. Farghaly and K. Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4):14:1–14:22, Dec. 2009.

[72] N. Farra, E. Challita, R. A. Assi, and H. Hajj. Sentence-level and document-level sentiment mining for Arabic texts. In *IEEE International Conference on Data Mining Workshops (ICDMW), 2010*, pages 1114–1119. IEEE, 2010.

[73] A. Farzindar and D. Inkpen. *Natural Language Processing for Social Media (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers, 2015.

[74] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[75] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[76] C. A. Ferguson. Diglossia. *word*, 15(2):325–340, 1959.

[77] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T. H. Kim. Mining twitter space for information: Classifying sentiments programmatically using java. In *Seventh International Conference on Digital Information Management (ICDIM), 2012*, pages 303–308. IEEE, 2012.

[78] A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, London, 4 edition, 2013.

[79] K. Fort, G. Adda, and K. B. Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, 2011.

[80] G. V. Glass and K. D. Hopkins. *Statistical methods in education and psychology*. Prentice-Hall Englewood Cliffs, NJ, 1970.

[81] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[82] P. C. Guerra, W. Meira Jr, and C. Cardie. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 443–452. ACM, 2014.

[83] N. Habash. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[84] N. Habash. Machine translation and Arabic language issues. In *The 10th biennial conference of the Association for Machine Translation in the Americas*

(AMTA-2012) San Diego, October 28, page 10. Center for Computational Learning Systems, Columbia University, 2012.

[85] N. Habash, M. T. Diab, and O. Rambow. Conventional orthography for dialectal Arabic. In *LREC*, pages 711–718, 2012.

[86] N. Habash, R. Eskander, and A. Hawwari. A morphological analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9. Association for Computational Linguistics, 2012.

[87] N. Habash and O. Rambow. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[88] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh. Morphological analysis and disambiguation for dialectal Arabic. In *HLT-NAACL*, pages 426–432, 2013.

[89] S. B. Hamouda and J. Akaichi. Social networks' text mining for sentiment classification: The case of facebook statuses updates in the Arabic spring era. *International Journal Application on Innovation in Engineering and Management*, 2(5):470–478, 2013.

[90] K. Hamza. Social media as a tool for social movements in Arab spring countries. In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, pages 71–74. ACM, 2014.

[91] A. Hatem and N. Omar. Syntactic reordering for Arabic-English phrase-based machine translation. In *Database Theory and Application, Bio-Science and Bio-Technology*, pages 198–206. Springer, 2010.

[92] L. Hong, G. Convertino, and E. H. Chi. Language matters in twitter: A large scale study. In *ICWSM*, 2011.

[93] F. Hoyt. Sentential negation marking in Palestinian and Moroccan Arabic: a study in comparative morpho-syntax. *Arabic Dialectology*, 2005.

[94] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. National Taiwan University, Taipei 106, Taiwan, 2003.

[95] P.-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[96] A. Htait, S. Fournier, , and P. Bellot. LSIS at SemEval-2016 Task 7: Using web search engines for English and Arabic unsupervised sentiment intensity prediction. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California, June 2016.

[97] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[98] H. Ibrahim, S. Abdou, and M. Gheith. MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis. In *IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), 2015*, pages 353–358, July 2015.

[99] H. S. Ibrahim, S. M. Abdou, and M. Gheith. Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis. *arXiv preprint arXiv:1506.01906*, 2015.

[100] M. M. Itani, L. Hamandi, R. N. Zantout, and I. Elkabani. Classifying sentiment in Arabic social networks: Naive search versus naive bayes. In *the 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), 2012*, pages 192–197. IEEE, 2012.

[101] M. Jarrar, N. Habash, D. Akra, and N. Zalmout. Building a corpus for Palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, 2014.

[102] T. Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[103] T. Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[104] T. Khalil, A. Halaby, M. Hammad, and S. R. El-Beltagy. Which configuration works best? an experimental study on supervised Arabic twitter sentiment analysis. In *in proceedings of the First Conference on Arabic Computational Linguistics (ACLing 2015) , co-located with CICLing 2015, Cairo, Egypt*, 2015.

[105] R. T. Khasawneh, H. A. Wahsheh, I. M. Alsmadi, and M. N. AI-Kabi. Arabic sentiment polarity identification using a hybrid approach. In *6th International Conference on Information and Communication Systems (ICICS)*, pages 148–153. IEEE, 2015.

[106] S. Kiritchenko, S. M. Mohammad, and M. Salameh. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California, June 2016.

[107] R. Kohavi. *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. PhD thesis, Stanford University, Department of Computer Science, Stanford University, 1995.

[108] M. Korayem, D. Crandall, and M. Abdul-Mageed. Subjectivity and sentiment analysis of Arabic: A survey. In *Advanced Machine Learning Technologies and Applications*, pages 128–139. Springer, 2012.

[109] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the OMG! *In Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11:538–541, 2011.

[110] A. D. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–290. ACM, 2010.

[111] F. Li, M. Huang, and X. Zhu. Sentiment analysis with global topics and local dependency. In *AAAI*, volume 10, pages 1371–1376, 2010.

[112] B. Liu. *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers, 2012.

[113] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[114] M. Marchetti-Bowick and N. Chambers. Learning for microblogs with distant supervision: Political forecasting with twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612. Association for Computational Linguistics, 2012.

[115] J. Martineau and T. Finin. Delta TFIDF: An improved feature space for sentiment analysis. In *ICWSM*, 2009.

[116] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pages 15–22, 2012.

[117] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *The semantic web: ESWC 2011 workshops*, pages 88–99. Springer, 2012.

[118] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[119] S. M. Mohammad, S. Kiritchenko, and X. Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 321–327, 2013.

[120] A. Mourad and K. Darwish. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. *WASSA 2013*, page 55, 2013.

[121] I. Mozeti, M. Grar, and J. Smailovi. Multilingual twitter sentiment classification: The role of human annotators. *PLoS ONE*, 11(5):1–26, 05 2016.

[122] M. Nabil, M. Aly, and A. Atiya. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[123] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[124] O. R. Nizar Habash and R. Roth. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009. The MEDAR Consortium.

[125] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[126] A. Osherenko. Towards semantic affect sensing in sentences. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, pages 41–44, 2008.

[127] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, 2010.

[128] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[129] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[130] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[131] A. Pasha, M. Al-Badrashiny, M. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[132] J. Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodssupport vector learning*, 3, 1999.

[133] R. Power, B. Robinson, and D. Ratcliffe. Finding fires with twitter. In *Proceedings of the Australasian Language Technology Association (ALTA) Workshop, Brisbane, Australia*, pages 80–89, 2013.

[134] M. Ptaszynski, R. Rzepka, K. Araki, and Y. Momouchi. Automatically annotating a five-billion-word corpus of japanese blogs for sentiment and affect analysis. *Computer Speech & Language*, 28(1):38–55, 2014.

[135] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491, Avignon, France, Apr. 2012. Association for Computational Linguistics.

[136] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[137] J. Read and J. Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 45–52. ACM, 2009.

[138] E. Refaee and V. Rieser. An Arabic twitter corpus for subjectivity and sentiment analysis. In *9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[139] E. Refaee and V. Rieser. Can we read emotions from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of Arabic twitter feeds. In *Proceedings of the the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data (ES3LOD)*, 2014.

[140] E. Refaee and V. Rieser. Evaluating distant supervision for subjectivity and sentiment analysis on Arabic twitter feeds. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014.

[141] E. Refaee and V. Rieser. Subjectivity and sentiment analysis of Arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)*, 2014.

[142] E. Refaee and V. Rieser. Benchmarking machine translated sentiment analysis for Arabic tweets. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–78, Denver, Colorado, June 2015. Association for Computational Linguistics.

[143] E. Refaee and V. Rieser. iLab-Edinburgh at SemEval-2016 Task 7: A hybrid approach for determining sentiment intensity of Arabic twitter phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California, June 2016.

[144] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[145] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov. SemEval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June 2015. Association for Computational Linguistics.

[146] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov. SemEval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[147] B. Roth, T. Barth, M. Wiegand, and D. Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78. ACM, 2013.

[148] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of ACL-08: HLT, Short Papers*, pages 117–120, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[149] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega. Bilingual experiments with an Arabic-English corpus for opinion mining. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 740–745, 2011.

[150] M. K. Saad and W. Ashour. Arabic morphological tools for text mining. *Corpora*, 18:19, 2010.

[151] D. Said, N. M. Wanas, N. M. Darwish, and N. Hegazy. A study of text preprocessing tools for Arabic text categorization. In *The second international conference on Arabic language*, pages 230–236, 2009.

[152] M. Saif and Z. Xiaodan. Sentiment analysis of social media texts. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP). Doha, Qatar. October 25-29, 2014*, page 10. Association for Computational Linguistics (ACL), 2014.

[153] M. Salameh, S. Mohammad, and S. Kiritchenko. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[154] W. Salloum and N. Habash. Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21. Association for Computational Linguistics, 2011.

[155] W. Salloum and N. Habash. Elissa: A dialectal to standard Arabic machine translation system. In *COLING (Demos)*, pages 385–392, 2012.

[156] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[157] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 1070–1079, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[158] B. Sharifi, M.-A. Hutton, and J. K. Kalita. Experiments in microblog summarization. In *IEEE Second International Conference on Social Computing (SocialCom)*, pages 49–56. IEEE, 2010.

[159] U. Shlonsky. *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax*, volume 11. Oxford University Press, 1997.

[160] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[161] Statistic-Brain-Research-Institute. Twitter Statistics. `http://www.statisticbrain.com/twitter-statistics/`. Accessed: 2015-10-03.

[162] C. Strapparava, O. Stock, and I. Alon. Corpus-based explorations of affective load differences in Arabic-Hebrew-English. In *COLING*, pages 1201–1208, 2012.

[163] J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer, 2013.

[164] M. Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2:325–347, 2016.

[165] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[166] Y.-j. Tang and H.-H. Chen. Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *LREC*, pages 1226–1229, 2012.

[167] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[168] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting

elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[169] UNESCO. World Arabic Language Day. `http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day/`. Accessed: 2014-05-06.

[170] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[171] H. Wang, V. R. Bommireddipalli, A. Hanafy, M. Bahgat, S. Noeman, and O. S. Emam. A system for extracting sentiment from large-scale Arabic social data. *arXiv preprint arXiv:1511.04661*, 2015.

[172] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[173] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[174] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

[175] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2013.

[176] A. Wolf. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3(5):827–833, 2000.

[177] J.-M. Xu, A. Bhargava, R. Nowak, and X. Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Machine Learning and Knowledge Discovery in Databases*, pages 644–659. Springer, 2012.

[178] K. Yeung. 61 languages are found on Twitter. Here is how they rank in popularity. `http://thenextweb.com/shareables/2013/12/10/61-languages-found-twitter-heres-rank-popularity/`. Accessed: 2014-03-15.

[179] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

[180] Z. Yuan and M. Purver. Predicting emotion labels for Chinese microblog texts. In *Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD)*, pages 40–47, Bristol, UK, Sept. 2012.

[181] W. Zaghouani. Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, pages 1–8, 2014.

[182] O. F. Zaidan. *Crowdsourcing Annotation for Machine Learning in Natural Language Processing Tasks*. PhD thesis, Johns Hopkins University, 2012.

[183] O. F. Zaidan and C. Callison-Burch. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 2*, pages 37–41. Association for Computational Linguistics, 2011.

[184] O. F. Zaidan and C. Callison-Burch. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202, 2014.

[185] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. F. Zaidan, and C. Callison-Burch. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59. Association for Computational Linguistics, 2012.

[186] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

[187] X. Zhu, S. Kiritchenko, and S. M. Mohammad. NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. *SemEval 2014*, 443, 2014.

[188] M. R. Zughoul. Diglossia in Arabic: investigating solutions. *Anthropological Linguistics*, pages 201–217, 1980.